Language Research and Development, Inc.

Erwin Tschirner, PhD
580 White Plains Rd., Ste. 660
Tarrytown, NY 10591
USA

May 31, 2020

Assessing Evidence of Validity and Reliability of the ACTFL Listening Proficiency Test (LPT)

Technical Report 2020/2-PUB-3

Prepared for:

American Council on the Teaching of Foreign Languages
Alexandria, VA

Prepared by:
Language Research and Development, Inc.

Dr. Erwin Tschirner
President

Table of Contents

ACTFL Listening Proficiency Test (LPT)

This evaluation of the ACTFL Listening Proficiency Test (LPT) follows the Examination Evaluation Checklist as provided by ACE. Where appropriate, the evaluation references documents provided as appendices. Item analysis results, reliability information, and evidence of validity are based on the following three languages: French, German, and Spanish.

1 General Information About the Examination

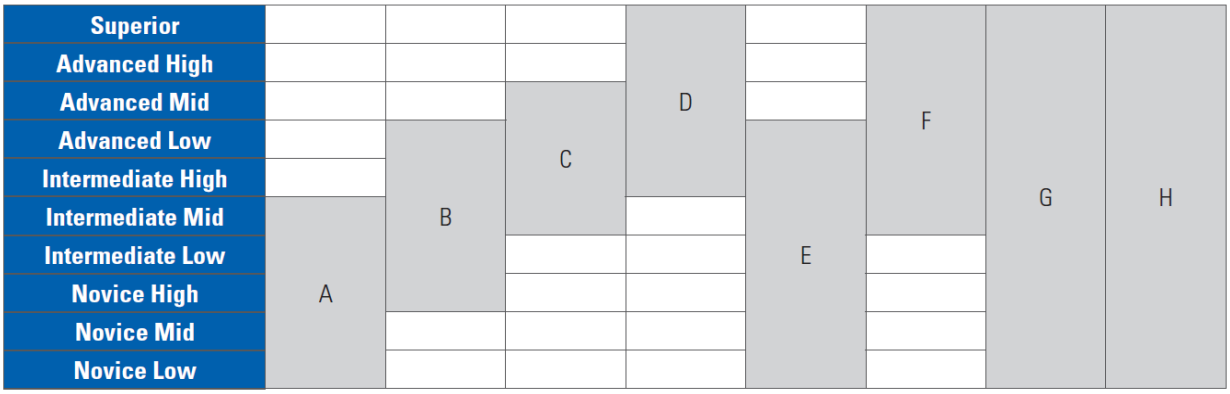
This section provides general information about the examination (see Appendix 1 – Familiarization Manual). The LPT is a standardized test for the global assessment of listening ability in a language. It is a carefully constructed assessment based on the *ACTFL Proficiency Guidelines 2012 – Listening* that evaluates Novice to Superior levels of listening ability. It is an online assessment. The test assesses specific ranges of proficiency. The available ranges are shown in Table 1 below. These options ensure that the test administered targets the range of the examinee's listening ability economically in terms of time and effort.

Sublevels and Number of Items per Test

There are five proficiency sublevels: Intermediate Low (IL); Intermediate Mid (IM); Advanced Low (AL); Advanced Mid (AM); and Superior (S). The number of tasks per test depends on the range of proficiency to be assessed (see Table 1 below). There are four two-sublevel (A-D), two three-sublevel (E-F), and two full-range tests (G-H). There are five listening passages per sublevel, each followed by three multiple-choice items (15 items per sublevel) with four options each, of which only one is correct. Version A includes five IL and five IM listening passages with 15 IL and 15 IM MC items for a total of 30 items; Version B includes five IM and five AL listening passages with 15 IM and 15 AL items for a total of 30 items; Version C includes five AL and five AM listening passages with 15 AL and 15 AM items for a total of 30 items; and Version D includes five AM and five S listening passages with 15 AM and 15 S items for a total of 30 items. Version E includes five IL, five IM, and five AL listening passages with 15 IL, 15 IM, and 15 AL items for a total of 45 items; Version F includes five AL, five AM, and five S listening passages with 15 AL, 15 AM, and 15 S items for a total of 45 items; and Version H includes five IL, five IM, five AL, five AM, and five S listening passages with 15 IL, 15 IM, 15 AL, 15 AM, and 15 S items for a total of 75 items. Version G is a semi-adaptive version of the test, which starts at Advanced Low, and moves to higher or lower level tasks based on the candidate's responses. Depending on the candidate's proficiency, it includes between 10 and 15 listening passages with 30 to 45 items. If the candidate is at least IM or at best AM, the test contains ten listening passages (five IM and five AL or five AL and five AM passages, respectively, for a total of 30 items). If the candidate is below IM or higher than AM, the test includes 15 listening passages (five passages each at IL, IM, and AL, or AL, AM, and S, respectively, for a total of 45 items). In addition to IL and IM, version A also assesses levels below IL, i.e. Novice Low (NL), Novice Mid (NH), and Novice High

(NH). The Novice levels are assessed on account of how much evidence there is of the Intermediate level, i.e. no or random (NL), emerging (NM), and developing but not sustained evidence (NH).

Table 1
Test Versions and Ranges Assessed



Item (Task) Types

There are four item types: Global, Detail, Selective, and Inference (see Table 2 for number of item types per sublevel):

- IL passages have one global, one selective, and one detail item.
- IM passages have one global and two detail items.
- AL passages have one global and two detail items.
- AM passages have one global, one detail, and one inference item.
- S passages have one global, one detail, and one inference item.

Table 2
Number of Item Types per Sublevel

Level	IL	IM	AL	AM	S
Number of items	Global: 5 Selective: 5 Detail: 5	Global: 5 Detail: 10	Global: 5 Detail: 10	Global: 5 Detail: 5 Inference: 5	Global: 5 Detail: 5 Inference: 5

Time Allotment

The time limit for a two-sublevel test is 50 minutes; for a three-sublevel test, it is 75 minutes; for the non-adaptive full-range test (H), it is 125 minutes, and for the semi-adaptive full-range test (G), it is 75 minutes. This amounts to approx. five minutes per task. However, there is only an overall time limit for the complete test. There is a time gauge to let examinees know how much time is still remaining.

2 Rationale and Purpose of the Examination

This section summarizes the rationale and purpose of the examination (see Appendix 3 – Design Statement and Appendix 11 – Examinee Handbook). The ACTFL LPT is the official listening proficiency test of the American Council on the Teaching of Foreign Languages (ACTFL). It assesses how well a person spontaneously reads listening passages in a world language when presented with passages and tasks as described in the *ACTFL Proficiency Guidelines 2012 – Listening* without access to dictionaries or grammar references. The *ACTFL Proficiency Guidelines 2012 – Listening* describe five major levels of proficiency: Distinguished, Superior, Advanced, Intermediate, and Novice. The description of each major level is representative of a specific range of abilities. Together these levels form a hierarchy in which each level subsumes all lower levels. The major levels Advanced, Intermediate, and Novice are divided into High, Mid, and Low sublevels. The ACTFL LPT assesses listeners' proficiency at all levels except Distinguished, i.e. from Novice Low to Superior.

3 Name(s) and Institutional Affiliations of the Principal Author(s) or Consultant(s)

The ACTFL LPT was developed by Dr. Erwin Tschirner (Gerhard Helbig Professor of German as a Foreign Language, University of Leipzig, and President of the Institute for Test Research and Test Development, Leipzig, Germany) and Dr. Olaf Bärenfänger (Director of the Language Learning Center, University of Leipzig, and Vice-President of the Institute for Test Research and Test Development, Leipzig, Germany). Dr. Tschirner and Dr. Bärenfänger also designed the item development process, and both are in charge of overall test validity and quality assurance.

Item development is managed by staff members of the Institute of Test Research and Test Development (ITT), including Jupp Möhring (M.A. in German as a Foreign Language, University of Leipzig), Elisabeth Muntschick (M.A. in German as a Foreign Language, University of Leipzig), and Robin Ide (M.A. in German as a Foreign Language, University of Leipzig).

Item Development Process

All items undergo a rigorous, standardized quality assured development process. Passage and item writers are native speakers of the language in question with a college degree in foreign language teaching or applied linguistics and with a considerable amount of language teaching and test writing experience. Test reviewers and senior test development officers are native or near-native speakers of the language in question and trained for language proficiency testing. Authors, reviewers, and final quality control specialists undergo a rigorous selection, training and certification process as well as ongoing quality assurance measures as appropriate for high stakes testing.

The training of test authors and reviewers constitutes an integral part of the Item Development Process. The Institute for Test Research and Test Development (ITT) regularly arranges item writing workshops consisting of several training sessions (one- and two-day workshops). The objective of the workshops is to train test authors and calibrate them with calibrated passages and items. The workshop facilitator is an ACTFL-trained and certified tester trainer. During these workshops, participants are familiarized with the Construct Matrix, the Item Writing Manual, and the Item Checklists while working individually and in groups. The workshop agenda includes the following activities: Sort the ACTFL Listening Proficiency Descriptors according to their proficiency levels; Complete the Construct Matrix with missing descriptors; Take an LPT to get familiar with the test; Get introduced to the Item Writing Manual and to Item Writing Do's and Don'ts; Get calibrated by benchmarking calibrated tests individually and in small groups; Write first drafts of items; and Take part in group discussions. After the workshop, there is a practice round and a certification round, in which participants author at least two passages and two sets of items at each sublevel, receive feedback on them, and get certified after passing the final review by a senior test development officer.

Items are developed in multiple stages in a controlled process. Certified authors who are native speakers of the target language develop passages and items according to the Item Writing Manual and the Construct Matrix and submit a first draft. The first draft is reviewed for style and correctness by another native speaker of the target language. The main focus of this review is to ensure that the passages are culturally and idiomatically authentic, well written, and able to hold the listener's interest. Tests are revised by the original author and submitted to an assessment specialist, who checks if the passages and items are at the appropriate levels, if the author has followed the instructions in the Item Writing Manual precisely, and if all items, keys, and distractors follow the norms established. This includes a first round of item sensitivity review to ensure that passages and items are not offensive or bias towards certain groups of examinees. The main focus is on the level appropriateness of the passages and the quality of the items. The assessment specialist is a native or near-native speaker of the target language. Tests are revised again by the original author or by a different native speaker author with similar qualifications.

Before the listening passages are recorded, they undergo a second round of sensitivity review. They are spoken by professional speakers who are trained to speak in a level-specific and criterion-based, authentic manner for the various proficiency levels and text types. The speakers are

experienced television or radio speakers, actors, speech scientists and/or world languages teachers with a substantial amount of teaching experience with a special talent for acting. They receive additional training for speaking LPT passages.

The sound recordings for the LPT are completed in sound studios, which comply with the guidelines and the specifications of public broadcasting (developed by the German television networks ARD and ZDF). Recordings and postproduction are undertaken by trained sound engineers. The recordings are professionally edited with background noise and other acoustic features that make them appear more authentic. After postproduction, several rounds of proof listening are carried out until the audio files are entered into the test system at LTI and checked again during User Assurance Testing (UAT).

The items are checked for spelling and punctuation and uploaded to the LTI Assessment System together with the audio files to begin UAT, which typically consists of two rounds and often results in additional revisions made to the items. The test then enters the operational testing phase with at least 300 examinees at all proficiency levels taking the test. Detailed data reports are developed using IRT analysis (Rasch modeling) (item difficulty logits, SEM, infit and outfit values, separation indices). Any misfitting items or any items that are too difficult or too easy for a particular level are revised or removed.

Table 3 lists the names, qualifications, and languages of most of the item writers and reviewers currently developing and reviewing items. The column *Other Languages* lists second languages with a proficiency of at least *Advanced Mid* but in many cases *Superior*. The columns *Assessment* and *Teaching* list the years of experience in both fields. People who show no experience in teaching in the table below are translators or interpreters, usually with a considerable amount of experience in their profession.

Table 3
Current Staff, Item Writers, and Reviewers

Name	Sex	Degree	Subject	Native Language	Other Languages	Assessment	Teaching
Erwin Tschirner	m	PhD (Berkeley)	Linguistics/SLA	German	Spanish	31	39
O. Bärenfänger	m	PhD (Bielefeld)	German	German	French	19	17
Jupp Möhring	m	MA (Leipzig)	German	German	English	9	12
Robin Ide	m	MA (Leipzig)	German	German	Spanish	7	7
Elisab. Muntschick	f	MA (Leipzig)	German	German	English	5	6
Elisa Hartmann	f	BA (Aachen)	German	German	English	2	3
Writer/Reviewer 1	f	MA (UFPR, Brasil)	Communication	Portuguese	German	15	20
Writer/Reviewer 2	f	MA (Leipzig)	German	Korean	German	4	5
Writer/Reviewer 3	f	MA (Leipzig)	Translation	French	German	4	
Writer/Reviewer 4	f	BA (Leipzig)	Communication	Arabic	German	1	1
Writer/Reviewer 5	f	BA (Leipzig)	Biology	Persian	German	1	1

Writer/Reviewer 6	f	MA (Fribourg)	German	German	French	7	8
Writer/Reviewer 7	f	MA (Chung-ang U)	Journalism	Korean	German	2	1
Writer/Reviewer 8	f	MA (Leipzig)	German	Italian	Spanish	1	2
Writer/Reviewer 9	f	PhD (Leipzig)	Media Studies	Chinese	German	8	10
Writer/Reviewer 10	f	BA (McGill)	Marketing	French	English	6	7
Writer/Reviewer 11	f	MA (U d´Orléans)	Comp Lit	French	German	5	37
Writer/Reviewer 12	m	MA (Isra U, Jordan)	Engineering	Arabic	German	3	1
Writer/Reviewer 13	f	MA (Guadalajara)	German	Spanish	German	3	7
Writer/Reviewer 14	f	BA (Wittenberg, OH)	Biology	English	German	2	9
Writer/Reviewer 15	m	MA (Guadalajara)	German	Spanish	German	3	3
Writer/Reviewer 16	f	PhD (Shahid Beheshti U)	German	Persian	German	1	1
Writer/Reviewer 17	f	MA (Teheran U)	German	Persian	German	12	12
Writer/Reviewer 18	m	MA (Guadalajara)	Philosophy	Spanish	German	10	14
Writer/Reviewer 19	f	BA (La Habana)	German/ESL	Spanish	German	2	4
Writer/Reviewer 20	f	BA (Leipzig)	Management	Russian	German	4	
Writer/Reviewer 21	f	BA (Leipzig)	German	Russian	German	2	
Writer/Reviewer 22	m	MA (Leipzig)	German	Spanish	German	4	4
Writer/Reviewer 23	f	MA (Leipzig)	European Studies	Russian	German	7	3
Writer/Reviewer 24	f	MA (Leipzig)	Spanish/Port.	Portuguese	Spanish	5	20

In addition to the people listed in Table 3, there were 27 additional people working in various consulting capacities, of which 19 were female and 8 were male.

4 Types of Scores Reported for Examinees

The ACTFL LPT is a proficiency test reporting proficiency levels as described in the *ACTFL Proficiency Guidelines 2012 – Listening*. Test scores are converted to ACTFL proficiency levels and reported as such (see Section 5 – Directions/Procedures for Scoring/Scoring Procedures/Keys).

In addition to the ACTFL listening proficiency level, the certificate also provides a brief description of what examinees who have reached a particular level can do. This helps examinees to place themselves within a continuum of proficiency levels (see Appendix 12 – Certificate).

5 Directions/Procedures for Scoring/Scoring Procedures/Keys

This section summarizes the scoring procedures (see Appendix 4 – Blueprint). The ACTFL LPT is machine-scored. At least two sublevels are administered and scored together, i.e. IL and IM, IM and AL, AL and AM, or AM and S. To assign a rating, the combined total of the two levels that

are rated is used. When there were more than two levels administered, the highest two levels that have at least 18 points between them are used. When there are no two levels that have a least 18 points between them, the highest two levels that have at least 11 points between them are used. When there are no two levels that have at least 11 points between them, the two lowest levels are used. Table 4 shows how test scores are converted to ACTFL ratings. (See Section 28 for information on how the cut scores were determined.)

Table 4
Scoring Algorithm

Sublevels Rated	Total Score	ACTFL Rating
IL-IM	0-11	NL
IL-IM	12-14	NM
IL-IM	15-17	NH
IL-IM	18-23	IL
IL-IM	24-30	IM
IM-AL	0-11	BR*
IM-AL	12-14	NH
IM-AL	15-17	IL
IM-AL	18-21	IM
IM-AL	22-23	IH
IM-AL	24-30	AL
AL-AM	0-11	BR*
AL-AM	12-14	IM
AL-AM	15-17	IH
AL-AM	18-23	AL
AL-AM	24-30	AM
AM-S	0-11	BR*
AM-S	12-14	IH
AM-S	15-17	AL
AM-S	18-21	AM
AM-S	22-23	AH
AM-S	24-30	S

*BR (Below Range) is assigned when the examinee's ability is lower than the lowest rating that may be assigned by a particular test version.

Table 4 shows what ratings are assigned to what scores given two particular sublevels. BR (Below Range) is assigned to scores of 0-11, because such scores could potentially be achieved by guessing only (see Section 28). For the sublevels IL and IM, the rating NL is assigned to scores of 0-11.

6 Specifications That Define the Domain(s) of Content, Skills, and/or Developed Abilities That the Exam Samples

This section summarizes the specifications that define the domain(s) of content, skills, and developed abilities that the exam samples (see Appendix 1 – Familiarization Manual, Appendix 2 – Assessment Use Argument, Appendix 3 – Design Statement, Appendix 4 – Blueprint, and Appendix 5 – Construct Matrix).

Based on the *ACTFL Proficiency Guidelines 2012 – Listening*, the construct matrix defines the domains of content, skills and abilities that the exam measures. The target language use (TLU) task that was selected as the basis for developing assessment tasks (passages and items) is listening in general, i.e. retrieving information from a variety of spoken passages in daily life, at work, university or school etc., indicating different aspects of comprehension (global, selective, detail understanding, or making inferences), depending on the sublevel. Tasks are described in terms of function, content, context, text type, vocabulary, grammar, and culture at all major ACTFL levels (see Table 5 for a summary of the task descriptors).

Table 5
Summary of Task Descriptors at the Proficiency Levels Represented by Test Tasks

	Function	**Content**	**Context**	**Text Type**	**Vocabulary**	**Grammar**	**Culture**
Superior	Argumentation; Supported opinion; Hypothesis	Familiar and unfamiliar abstract topics	Professional; Academic; Literary	Complex, lengthy passages	Broad; Precise; Specialized	Complex structures	Cultural references; Aesthetic properties
Advanced	Description; Narration; Exposition; Explanation;	Concrete current and general interest topics	Public; Education; Work; News	Paragraph-based connected passages with a clear predictable structure	Broad general vocabulary	Sequencing; Time frames; Chronology	Most common cultural patterns
Intermediate	Convey basic information	Highly familiar everyday content	Highly familiar everyday contexts	Simple, predictable, loosely connected passages	High frequency vocabulary	Simple sentence patterns and strings of sentences	Some of the most common cultural patterns

- The term *function* refers to the different purposes spoken passages may have such as instruction, description, narration, explanation, or argumentation.

- The term *content* refers to the general content areas that the listener can understand in the language.
- The term *context* refers to the different domains in which discourse occurs such as the public, educational or work domain.
- The term *text type* refers to the quantity, quality and organization of passages that the listener can understand in the language.
- The term *vocabulary* refers to the range of vocabulary the listener can understand in the language.
- The term *grammar* refers to the range of grammatical structures that the listener is able to use for comprehension purposes.
- The term *culture* refers to the range of idiomatic expressions and cultural references the listener can understand in the language.

7 Statement of the Exam's Emphasis on Each of the Content, Skill, and/or Ability Areas

The contents, skills and ability areas are based on the *ACTFL Proficiency Guidelines 2012 – Listening*. Each exam contains items for at least two sublevels. Thus, at least ten passages and 30 items form the basis of a rating. This allows the test to assess a representative sample of real-life topics and to make a meaningful statement about the language proficiency of an examinee. Depending on the sublevels assessed, the listening passages have different functions such as description, narration, explanation, exposition, argumentation, and hypothesis and different contexts such as familiar everyday contexts, work, public, education, academic, professional and art contexts. For example, the test that assesses the sublevels Advanced Mid and Superior contains ten passages, which represent the functions of both levels, i.e. description, narration, explanation, and exposition at the Advanced level and argumentation, supported opinion, and hypothesis at the Superior level. A similar distribution applies to content and genre. The test involves passages of concrete, current, and general interest topics as well as familiar and unfamiliar abstract topics such as discussion between educated native speakers, radio broadcasts, news stories, oral reports, and lectures concerned with contemporary social problems, biographical accounts, stories, and opinion/editorial pieces, analyses and commentaries.

8 Rationale for the Kinds of Tasks (Passages and Items) Included in The Exam

This section presents the rationale for the kinds of items included in the exam (see Appendix 3 – Design Statement, Appendix 4 – Blueprint, and Appendix 5 – Construct Matrix). Please see sections 6 and 7 for the rationale for the kinds of passages included in the exam. This is the rationale for the items:

There are four item types: Global (for the sublevels IM to S), Detail (for all sublevels), Selective (for IL only), and Inference (for the sublevels AM to S). These item types were derived from the

ACTFL Proficiency Guidelines 2012 – Listening and from the cognitive processing approach to defining comprehension of Weir and Khalifa (2008). Their model was developed for reading comprehension but it applies equally well for listening comprehension, in particular, their model of listener intent (goal setter) with its dimensions of local (detail) vs. global (gist) and careful vs. expeditious listening. *Expeditious* was redefined as *casual* for the model used by the LPT. The distribution of item types across sublevels is as follows:

- IL passages have one selective and two detail items.
- IM passages have one global and two detail items.
- AL passages have one global and two detail items.
- AM passages have one global, one detail, and one inference item.
- S passages have one global, one detail, and one inference item.

Passages and items align with each other with respect to function. Intermediate passages, e.g., may be understood sentence by sentence. Intermediate items consequently focus on information contained within the context of an individual sentence-length utterance. Advanced passages consist of descriptive and narrative passages that require paragraph-length comprehension and the understanding of cohesive devices to signal, e.g., sequencing, time frames, and chronology. Advanced items consequently focus on information that is distributed across several sentence-length utterances within a passage. Depending on the sublevel, item types, therefore, are defined differently as follows:

<u>Global</u>

- IL: Able to identify general subject matter, gets an idea of the content. The general subject matter is put in very broad terms. Distractors are viable passage-based options, i.e. there are words and phrases in the passage that refer plausibly to these options.
- IM: Able to identify general subject matter, understands the gist of the passage. The general subject matter is put in terms that require a global understanding of the passage at hand.
- AL: Ability to understand the main idea depends on comprehending supporting details. Examinee needs to understand some details to answer the question correctly. The correct answer needs to be synthesized from understanding different parts of the passage. The main idea is of a factual nature rather than focusing on author intent.
- AM: Ability to understand the main idea and/or argument depends on comprehending supporting details. The correct answer is spread out over different parts of the passage. It is based on what the speaker or speakers intended to say. Speaker intent is clearly signaled.
- S: Fully able to understand the main argument and all supporting facts. It is the main argument the speaker or speakers are making. The correct answer is spread out over different parts of the passage. Distractors refer to other arguments the speaker or speakers are making or to an argument they could be making based on statements contained in the passage.

Detail

- IL: Able to understand simple single facts. These facts are the easiest to understand aurally and do not necessarily have to be important for the passage as a whole. Distractors must be viable passage-based options, which must be clearly false.
- IM: Able to understand single straightforward facts. These facts contribute to the gist of the passage. Still, their comprehension only requires understanding single simple sentence-length utterances. Distractors must be viable passage-based options. Key must use synonyms or paraphrases that consist of highly frequent or shared international vocabulary pronounced similarly in both languages.
- AL: Able to understand explicitly mentioned facts and thoughts. They go beyond simple sentence-based facts. Their understanding is dependent on understanding the gist of the passage. They require understanding more than one sentence-length utterance. Distractors focus on other relevant facts mentioned in the passage. Key must use synonyms or paraphrases that contain general vocabulary.
- AM: Able to understand explicitly mentioned facts, thoughts, and argument. Their understanding is dependent on understanding the gist of the passage. They require understanding complete subsections of the passage rather than single sentences. Keys and distractors focus on explicitly mentioned facts or argument. Key must use synonyms and paraphrases that contain a broad general vocabulary.
- S: Able to understand argument, finer points of detail and abstraction. They require understanding complete subsections of the passage rather than single sentences. Keys and distractors focus on finer points of detail and abstraction that support the main argument of the passage. Key must use synonyms and paraphrases. Stem, key, and distractors commonly contain precise, specialized and low-frequency vocabulary.

<u>Selective</u>

- IL: Able to understand familiar words and very basic phrases. Both stem and options repeat words and phrases from the passage. The main task is to understand the question and to notice the answer in the passage. Both key and distractors need to contain language that is taken from the passage.

<u>Inference</u>

- AM: Able to identify the main conclusions in clearly signaled explanatory or argumentative passages and to make straightforward inferences. Items refer to the complete passage and focus on something that is clearly understood but not explicitly mentioned in the passage.
- S: Able to infer attitude, mood, and intentions; able to infer implied as well as stated opinions; able to draw conclusions. Items refer to the complete passage, the main argument or subordinate arguments. They refer to something the speaker or speakers clearly had in mind, to their attitude towards the issue, or the reasons why they said what they said.

Item Difficulty

Items align with their level with respect to function, vocabulary, and grammar.

- IL: Most frequent common basic words and phrases, common names, cognates and shared international vocabulary pronounced similarly; short, simple sentence-length utterances, predominantly in the present tense.
- IM: High-frequency words and phrases, cognates, and shared international vocabulary; short simple sentence-length utterances.
- AL: Variety of frequent words and phrases, cognates, and shared international vocabulary; longer and more complex turns containing some subordinate clauses, prepositional phrases and other features of connected discourse.
- AM: Broad active listening vocabulary and some low-frequency words and expressions; complex paragraph-length turns containing subordinate clauses, prepositional phrases and other features of connected discourse.
- S: Precise, often specialized and low-frequency vocabulary and expressions, including idioms and colloquialisms; complex paragraph-length turns containing subordinate and prepositional clauses, gerunds and participial clauses referring to complex, abstract, and hypothetical argumentation and relationships.

9 Information About Why Each Task is Included in the Test and Information About the Adequacy of the Tasks on the Exam as a Sample from the Domain(s)

The ACTFL LPT includes a broad spectrum of genres and topic categories to assure that the test adheres to its construct and consists of topics and language that are relevant for examinees. Each topic is used only once at any one level to provide a representative sample of the language proficiency of examinees across a broad range of topics. Tables 6 and 7 below provide an example of the genres and topics included in a test. Note that these are open lists that continue to be updated.

Table 6
Task Genres per Sublevel

IL	IM	AL	AM	S
Simple Announcements	Simple Announcements			
Simple Conversations	Simple Conversations			
Short routine telephone or online conversations	Short routine telephone or online conversations			

		Interviews	Interviews	Interviews
		News Items	News Items	News Items
		Narratives	Narratives	Narratives
		Oral Reports	Oral Reports	Oral Reports
			Opinion Pieces	Opinion Pieces
			Short Lectures	Short Lectures
			Debates	Debates
			Technical Discus-sions	Technical Discus-sions

Table 7
Task Topics and Subtopics

Topics	**Subtopics**
Arts	Age
Business & Commerce	Airport
Daily Life	Animals
Education	Brain
Family	Children
Fiction	Cinema
Food	College
Free time	Computer
Government and Politics	Directions
Health & Wellbeing	Drugs
Home	Environment
Law & Crime	Gender
Nature	History
News	Hobbies
Popular culture	Hospital
Science	Hotel
Society	Internet
Sports	Interview
Style	Languages
Technology	Literature
Travel	Living
Work	Love
	Math
	Meeting
	Money
	Moving
	Museum
	Music
	New Job

	People
	Pets
	Plans
	Plants
	Problems
	Recipe
	Religion
	Restaurant
	Routine
	School
	Shopping
	Souvenirs
	Theater
	Trade
	Tradition
	Traffic
	Train
	Transportation
	Trends
	Trips
	TV
	Weather

Subtopics may be subtopics of more than one main topic. Each exam provides a representative sample of the construct by including a broad spectrum of topics, subtopics, genres, and rhetorical organization (text type). The LPT is commonly taken as a two-sublevel test and consists of ten passages, five at each level. The ten passages are chosen to provide a representative statement of the language proficiency of the examinee. In the following, three different examples of two-level tests are presented to show how the passages reflect the *ACTFL Proficiency Guidelines 2012 – Listening* and how the test ensures selecting a diverse and representative sample of the topics, subtopics, genres, and rhetorical organization of passages listeners are able to read at each level.

Example 1 represents a test that spans the sublevels NL to IM. Passages and items are at the sublevels IL and IM. NH is defined as responding correctly to 50% of the Intermediate items, NM responding correctly to 40% of the items, and NL to less than 40%. Passage topics, subtopics, genres and rhetorical organization are based on the ACTFL level descriptions as follows:

Intermediate Low

At the Intermediate Low sublevel, listeners are able to understand some information from sentence-length speech, one utterance at a time, in basic personal and social contexts, though

comprehension is often uneven. At the Intermediate Low sublevel, listeners show little or no comprehension of oral texts typically understood by Advanced-level listeners.

Intermediate Mid

At the Intermediate Mid sublevel, listeners are able to understand simple, sentence-length speech, one utterance at a time, in a variety of basic personal and social contexts. Comprehension is most often accurate with highly familiar and predictable topics although a few misunderstandings may occur. Intermediate Mid listeners may get some meaning from oral texts typically understood by Advanced-level listeners.

Table 8 shows the variety and distribution of topics, subtopics, genres and rhetorical organization in a typical NL to IM test.

Table 8
Distribution of Topics, Subtopics, Genres, and Rhetorical Organization in a Typical NL to IM Test

Passage	Topic	Subtopic	Genre	Rhetorical Organization
IL.1	Free Time	Shopping	Announcement	Instruction
IL.2	Food	Restaurant	Simple Conversation	Simple Description
IL.3	Family	People	Telephone Conversation	Simple Description
IL.4	Daily Life	Pets	Simple Conversation	Instruction
IL.5	Arts	Theater	Announcement	Simple Description
IM.1	Daily Life	Routine	Simple Conversation	Simple Description
IM.2	Sports	Plans	Virtual Exchange	Simple Narration
IM.3	Daily Life	Moving	Simple Conversation	Simple Narration
IM.4	Work	Routine	Simple Conversation	Simple Narration
IM.5	Society	Literature	Announcement	Simple Description
Distribution	3x Daily Life 1x Free Time 1x Food 1x Family 1x Arts 1x Sports 1x Work 1x Society	1x Shopping 1x Restaurant 1x People 1x Pets 1x Theater 2x Routine 1x Plans 1x Moving 1x Literature	3x Announcement 5x Simple Conversation 1x Telephone Conversation 1x Virtual Exchange	2x Instruction 5x Description 3x Narration

Example 2 represents a test that spans the sublevels IM to AM. Passages and items are at the levels AL and AM. IH is defined as responding correctly to 50% of the Advanced items, and IM as responding correctly to 40% of the items. Responding to less than 40% of the items correctly is defined as Below Range (BR), i.e. as below the lowest sublevel the test is able to assess reliably.

Passage topics, subtopics, genres and rhetorical organization are based on the ACTFL level descriptions as follows:

Advanced Low

At the Advanced Low sublevel, listeners are able to understand short conventional narrative and descriptive passages with a clear underlying structure though their comprehension may be uneven. These passages predominantly contain high-frequency vocabulary and structures. Listeners understand the main ideas and some supporting details. Comprehension may often derive primarily from situational and subject-matter knowledge. Listeners at this level will be challenged to comprehend more complex passages.

Advanced Mid

At the Advanced Mid sublevel, listeners are able to understand conventional narrative and descriptive passages, such as expanded descriptions of persons, places, and things and narrations about past, present, and future events. The speech is predominantly in familiar target-language patterns. Listeners understand the main facts and many supporting details. Comprehension derives not only from situational and subject-matter knowledge, but also from an increasing overall facility with the language itself. Listeners at this level may derive some meaning from passages that are structurally and/or conceptually more complex.

Table 9 shows the variety and distribution of topics, subtopics, genres and rhetorical organization in a typical IM to AM test.

Table 9
Distribution of Topics, Subtopics, Genres, and Rhetorical Organization in a Typical IM to AM Test

Task	Topic	Subtopic	Genre	Rhetorical Organization
AL.1	Society	Trends	Simple Story	Narration
AL.2	Daily Life	People	Interview	Description
AL.3	Work	Children	Simple Story	Narration
AL.4	Travel	Money	New Item	Explanation
AL.5	Travel	Trips	Oral Report	Description
AM.1	Society	People	News Item	Narration
AM.2	Education	School	Story	Narration
AM.3	Government/Politics	Plans	Short Lecture	Explanation
AM.4	Arts	Cinema	Interview	Explanation
AM.5	Society	Tradition	Interview	Exposition
Distribution	3x Society 1x Daily Life 1x Work 2x Travel 1x Education	1x Trends 2x People 1x Children 1x Money 1x Trips	3x Story 3x Interview 2x News Item 1x Oral Report 1x Short Lecture	4x Narration 2x Description 3x Explanation 1x Exposition

	1x Government and politics 1x Arts	1x School 1x Plans 1x Cinema 1x Tradition		

Example 3 represents a test that spans the sublevels IH to S. Passages and items are at the levels AM and S. AL is defined as responding correctly to 50% of the AM and S items, and IH as responding correctly to 40% of the items. Responding to less than 40% of the items correctly is defined as Below Range (BR), i.e. as below the lowest sublevel the test is able to assess reliably. Passage topics, subtopics, genres and rhetorical organization are based on the ACTFL level descriptions as follows:

Advanced Mid

At the Advanced Mid sublevel, listeners are able to understand conventional narrative and descriptive passages, such as expanded descriptions of persons, places, and things and narrations about past, present, and future events. The speech is predominantly in familiar target-language patterns. Listeners understand the main facts and many supporting details. Comprehension derives not only from situational and subject-matter knowledge, but also from an increasing overall facility with the language itself. Listeners at this level may derive some meaning from passages that are structurally and/or conceptually more complex.

Superior

At the Superior level, listeners are able to understand speech in a standard dialect on a wide range of familiar and less familiar topics. They can follow linguistically complex extended discourse such as that found in academic and professional settings, lectures, speeches, and reports. Comprehension is no longer limited to the listener's familiarity with subject matter, but also comes from a command of the language that is supported by a broad vocabulary, an understanding of more complex structures and linguistic experience within the target culture. Listeners at the Superior level can understand not only what is said, but sometimes what is left unsaid; that is, they can make inferences.

Superior-level listeners understand speech that typically uses precise, specialized vocabulary and complex grammatical structures. This speech often deals abstractly with topics in a way that is appropriate for academic and professional audiences. It can be reasoned and can contain cultural references.

Table 10
Distribution of Topics, Subtopics, Genres, and Rhetorical Organization in a typical IH to S test

Task	Topic	Subtopic	Genre	Rhetorical Organization
AM.1	Society	People	News Item	Narration
AM.2	Education	School	Story	Narration

AM.3	Government and Politics	Plans	Short Lecture	Explanation
AM.4	Arts	Cinema	Interview	Explanation
AM.5	Society	Tradition	Interview	Exposition
S.1	Business & Commerce	Money	Debate	Hypothesis
S.2	Government and Politics	Reform	Short Lecture	Argument
S.3	Food	Trends	News Item	Narration
S.4	Technology	Reform	Opinion Piece	Hypothesis
S.5	Science	Problems	Debate	Argument
Distribution	2x Society 1x Education 2x Government and Politics 1x Arts 1x Business & Commerce 1x Food 1x Technology 1x Science	1x People 1x School 1x Plans 1x Cinema 1x Tradition 1x Money 2x Reform 1x Trends 1x Problems	2x News Item 1x Story 2x Short Lecture 2x Interview 2x Debate 1x Opinion Piece	3x Narration 2x Explanation 1x Exposition 2x Hypothesis 2x Argument

As these examples show, the tasks in any single exam cover a broad spectrum of topics, subtopics, genres, and rhetorical organization to provide a solid and representative statement of the listening proficiency of examinees. The qualitative and quantitative analyses of the three representative test ranges also provides evidence that the test items represent the domains of knowledge and abilities the test claims it does well. We will refer to this section again in the section on content validity below (Section 18).

10 Information About the Currency and Representativeness of the Test's Items

The *ACTFL Proficiency Guidelines 2012 – Listening* represent the current state of knowledge of second language listening proficiency at various levels of proficiency. Section 7 showed that the number and distribution of topics, subtopics, genres, and rhetorical organization are representative of the proficiency levels identified by the *ACTFL Proficiency Guidelines 2012 – Listening* and Section 6 showed the same for the items (listening goals), completing the categories used by ACTFL in its level descriptions.

11 Description of the Item Sensitivity Panel Review

The main sensitivity concerns in second language testing include the kinds of topics and the language used. Neither topics nor language should be offensive toward any examinees. Item writers are instructed to avoid topics such as drugs, sexuality, war, violence, etc. that may engender strong emotional reactions as well as discriminating and linguistically inappropriate content to ensure equal access to the passages for all examinees. It is neither economically feasible nor, indeed, necessary to use panels instead of individual reviewers to review the appropriateness of

topics, the situations described, the arguments provided, and the language used in world languages tests. Item sensitivity review is included in all phases of the item development process: the writing, the revisions, and the quality assurance phase (UAT). In addition, it is part of the item revision process after IRT analyses have been completed, again in multiple stages. Within the life cycle of a test form, item sensitivity review is part of the following stages.

1. Test writers are instructed to ensure that the content and language of all passages and items are appropriate.
2. Item reviewers are instructed to flag inappropriate content and/or language. There are two reviews completed by two different reviewers, one focusing on content and style and the other one focusing on level appropriateness and item quality.
3. Before the listening passages are spoken and recorded, a quality assurance reviewer is instructed to complete a final round of sensitivity review.
4. In the revision cycle after the IRT analysis, test writers revising flagged items are instructed to read all passages and items, not only flagged ones, to ensure the appropriateness of content and language, in addition to flagging outdated content.
5. Revised passages and/or items are reviewed by item specialists who also inspect the appropriateness of the content and the language of the revised passages or items.

During the first cycle (steps 1-3), therefore, the appropriateness of content and language is checked by four different people, and during the second cycle (steps 4-5) by an additional two different people for a total of six different people altogether. Great care is taken to ensure an equal distribution of male and female item writers, item reviewers, and quality assurance reviewers as well as of people of various ethnic, social, and regional backgrounds and sexual orientation.

12 Information About Whether and/or How the Items Were Pretested Before Inclusion into the Final Form

All forms go through a rigorous development process (see Item Development Process in Section 3). There is no pretesting. IRT analyses are performed, generally, after approx. 300-400 test administrations, after which some items are, generally, revised. To date, all reports have found that each released form showed good psychometric properties with high overall Rasch separation reliability to meet the requirements of a high-stakes test (see Section 13 below).

13 Item Analysis Results (e.g., Item Difficulty, Discrimination, Correlation with External Criteria)

This section presents the item analysis results (e.g., item difficulty, discrimination, correlation with external criteria) (see Appendix 6 – Technical Report and Appendix 7 – French, German, and Spanish LPT Data Reports).

Data reports are completed for all test forms, generally, after 300-400 test administrations when sufficient numbers of examinees have taken each individual sublevel (IL, IM, AL, AM, S). Data reports provide the date on which the report was completed, the name of the test, e.g., Spanish LPT 01, the name of the person completing the report, and the number of participants. The data are analyzed using item response theory (IRT). The IRT model used is the Rasch model for dichotomous items.

For each item, the data reports provide the number of cases; the item difficulty (measure) reported in logits; the standard error of the mean (SEM) also reported in logits; infit and outfit statistics; and the separation index (point-biserial item-scale correlations) to indicate how well the item discriminates between examinees at various proficiency levels. A comment column completes the item table. In addition, the data reports provide the overall separation reliability and overall model fit, and they make recommendations with respect to item difficulty, separation, overall reliability, and construct validity. They also list the nine anchor items and indicate from which form they were derived and their IDs. Each report concludes with a general statement as to the quality of the psychometric properties of the test and its usability for high stakes testing.

The item difficulty measure of an item expressed in logits should fall within a particular range for each sublevel. These ranges vary from language to language. If the item difficulty falls outside the range of the sublevel but stays within the range of an adjoining sublevel, the item is flagged for inspection (yellow). If it also falls outside the range of an adjoining sublevel, it is flagged for revision (red).

Fit statistics indicate the degree to which a test item meets the Rasch model expectations. Fit values between .5 and 1.5 mean-squares are the most productive values for measurement. Fit values between 0 and 0.5 as well as 1.5 and 2.0 mean-squares are unproductive but not degrading. Fit values larger than 2.0 mean-squares indicate too much variance, degrading the measurement. Whereas infit statistics are sensitive to the competence range for which the test was designed, outfit statistics are sensitive to outliers. Traditionally, infit statistics are considered more important than outfit statistics. Items with infit values above 2.0 are recommended for revision and flagged red. Items with outfit values above 2.0 are recommended for inspection and flagged yellow.

Separation indices should not fall below 0.20. Unlike the Rasch item difficulty estimates, item-scale correlations are sample-dependent. Sampling errors such as participants being more or less proficient than expected, affect the item discrimination parameter. Items with separation indices between 15 and 19 are flagged for inspection (yellow), while items with separation indices below 15 are flagged for revision (red).

The comment column spells out the action recommended (inspection or revision) and the main reason(s) such as inappropriate item difficulty, infit statistic, or separation index.

The overall separation reliability should not be lower than 0.8, and the overall model fit statistics should ideally be between 0.5 and 1.5 but values below 0.5 and between 1.5 and 2.0 are acceptable. Good overall and item infit statistics, moreover, provide evidence of construct validity because they indicate that the test form measures the proficiency range for which it was designed (see also Section 20).

Results

The item difficulty and discrimination parameters for the LPT are presented for the three selected languages, i.e. French, German and Spanish. The results of the most recent forms are included below, i.e. French 02, German 02 and Spanish 03 (see Appendix 7 – French, German, and Spanish LPT Data Reports for all French, German, and Spanish forms).

The item difficulty measure is reported in logits as estimated by the Rasch model for dichotomous items (see Tables 11-13). Probabilistic test theory (Rasch model) yields information that is sample-independent and expresses item difficulty across all proficiency levels on the same metric. The standard error of measurement (SEM) of the difficulty estimate is also reported in logits. Please note that these difficulty parameters cannot be compared directly across languages.

Tables 11-13 show a variety of measures for all of the items in the test. The items are listed in rows. They are coded by level, task, and item. A1 indicates IL, A2 indicates IM, B1 indicates AL, B2 indicates AM, and C1 indicates Superior. The first digit after the sublevel indicates the listening passage, i.e. passages 1 through 5, and the second digit after the sublevel indicates the item, i.e. items 1 through 3. Thus, A1.1.1 indicates IL listening passage 1 item 1.

Column 2 provides the number of examinees (*N*) responding to a particular item; column 3 provides the item difficulty (measure) in logits; and column 4 the standard error of measurement (SEM), also expressed in logits. Columns 5 and 6 provide the Rasch infit and outfit values in mean-squares (MNSQ). Column 7 provides the separation index (item discrimination) expressed as a point-biserial correlation (r_{pb}); and Column 8 provides the action recommended together with the main reason(s). For each language, the mean difficulty logic of all items is set to 0.

Conspicuous items requiring action are flagged. A yellow flag means that the item needs to be inspected, and revised if needed, while a red flag means that the item needs to be revised. Items with difficulty measures one standard deviation (SD) below or above the mean of the sublevel are flagged yellow when the measure falls within the one SD ranges of an adjoining sublevel and red when the measure falls outside the one SD ranges of the adjoining sublevel. Infit values above 2.0 MNSQ are flagged red and outfit values above 2.0 MNSQ are flagged yellow. Separation indices below 0.15 are flagged red and indices between 0.15 and 0.19 are flagged yellow. Table 11 provides the item characteristics for French LPT 02.

Table 11
Item Characteristics French LPT 02

Item	Number of Cases	Measure	SEM	Infit	Outfit	Separation Index	Comment
A1.1.1	464	-4.23	.20	.82	.36	.42	Item may be too easy. Inspect.
A1.1.2	464	-3.14	.14	.88	.67	.45	
A1.1.3	464	-1.97	.11	1.05	1.06	.41	
A1.2.1	464	-3.42	.15	.94	.81	.38	
A1.2.2	464	-3.05	.14	.87	.76	.46	
A1.2.3	464	-4.27	.20	.81	.32	.42	Item may be too easy. Inspect.
A1.3.1	464	-2.31	.12	.91	.80	.50	
A1.3.2	464	-.93	.11	.85	.83	.58	
A1.3.3	464	-.52	.11	.99	1.12	.46	
A1.4.1	464	-2.29	.12	.91	.85	.49	
A1.4.2	464	-1.48	.11	.89	.85	.55	
A1.4.3	464	-2.60	.12	.88	.69	.51	
A1.5.1	463	-2.16	.11	.99	1.02	.43	
A1.5.2	462	-3.04	.14	.86	1.16	.44	
A1.5.3	462	-1.92	.11	1.07	1.06	.40	
A2.1.1	798	.51	.09	.89	1.00	.48	
A2.1.2	798	-1.25	.08	.92	.87	.52	
A2.1.3	798	-1.30	.08	.86	.77	.58	
A2.2.1	797	-.99	.08	1.13	1.22	.35	
A2.2.2	795	-2.01	.09	.82	.69	.58	
A2.2.3	798	.62	.09	1.38	1.98	.04	Separation index below threshold. Revise.
A2.3.1	797	-1.43	.08	.90	.88	.53	
A2.3.2	798	-1.08	.08	.88	.85	.55	
A2.3.3	798	-.28	.08	.83	.84	.58	
A2.4.1	798	-.69A	.08	1.43	1.60	.12	Separation index okay in LPT 01. Ignore.
A2.4.2	795	-.52A	.08	1.04	1.11	.42	
A2.4.3	797	.41A	.09	.96	1.25	.38	
A2.5.1	798	-1.57	.09	1.11	1.28	.35	
A2.5.2	798	-.58	.08	1.04	1.07	.42	
A2.5.3	798	-1.97	.09	.80	.67	.59	
B1.1.1	870	-.43	.09	.95	.75	.48	
B1.1.2	871	.25	.08	.99	.96	.46	
B1.1.3	868	.31	.08	.89	.82	.54	
B1.2.1	870	.77	.08	.98	1.04	.46	
B1.2.2	869	.91	.08	.94	.99	.50	
B1.2.3	868	-.56	.10	.98	.85	.44	
B1.3.1	869	-.59	.10	.84	.68	.54	
B1.3.2	869	-.39	.09	.85	.68	.55	
B1.3.3	868	-.30	.09	.88	.79	.52	
B1.4.1	870	-.28A	.09	.94	.81	.37	
B1.4.2	870	.71A	.08	.95	.95	.51	
B1.4.3	864	1.24A	.08	.86	.88	.56	
B1.5.1	872	.66	.08	.95	.91	.50	

B1.5.2	870	.32	.08	.87	.90	.55	
B1.5.3	871	.10	.08	.89	.82	.53	
B2.1.1	652	.85	.09	1.11	1.16	.25	
B2.1.2	652	.76	.09	.80	.71	.58	
B2.1.3	652	2.30	.09	.96	1.01	.39	
B2.2.1	652	2.52	.09	1.15	1.43	.15	Separation index below threshold. Inspect.
B2.2.2	652	2.81	.09	1.24	1.42	.07	Separation index below threshold. Revise.
B2.2.3	651	2.97	.10	.99	1.09	.32	Item too difficult. Revise.
B2.3.1	652	3.20	.10	1.07	1.26	.20	Item too difficult. Revise.
B2.3.2	650	.65	.09	.99	.99	.38	
B2.3.3	652	.36	.10	1.10	1.16	.24	
B2.4.1	652	1.30A	.09	1.02	1.03	.36	
B2.4.2	651	1.34A	.09	1.10	1.09	.35	
B2.4.3	651	.30A	.10	1.01	.94	.42	
B2.5.1	652	1.99	.09	.98	.95	.40	
B2.5.2	652	.39	.10	1.01	.98	.34	
B2.5.3	652	1.27	.09	.88	.85	.51	
C1.1.1	608	.69	.09	1.01	.98	.36	
C1.1.2	608	1.14	.09	1.01	1.02	.36	
C1.1.3	604	3.08	.10	.93	.95	.38	
C1.2.1	608	2.37	.09	1.17	1.27	.18	Separation index below threshold. Inspect.
C1.2.2	607	3.15	.10	1.04	1.31	.24	
C1.2.3	608	1.09	.09	1.00	1.01	.37	
C1.3.1	609	2.40	.09	1.18	1.38	.14	Separation index below threshold. Revise.
C1.3.2	608	2.34	.09	1.00	1.13	.34	
C1.3.3	609	2.21	.09	.96	.99	.40	
C1.4.1	609	.28	.10	.99	.98	.35	Item may be too easy. Inspect.
C1.4.2	610	.96	.09	1.06	1.05	.31	
C1.4.3	607	2.21	.09	1.15	1.26	.20	
C1.5.1	608	2.08	.09	1.13	1.18	.23	
C1.5.2	607	1.59	.09	.98	.98	.40	
C1.5.3	600	2.53	.09	1.03	1.09	.31	

Table 11 shows the item characteristics for French LPT 02. It shows that the overall item difficulty increased with the sublevels tested as expected. The precision of the item difficulty parameter was high, as suggested by the SEM, varying from .10 to .20 at the IL level, from .08 to .09 at the IM level, from .05 to .10 at the AL level, from .09 to .10 at the AM level, and from .09 to .10 at the Superior level. All infit values were between 0.5 and 1.5 and many of them were close to 1.0, indicating that the items fit the model well. The great majority of the outfit values also ranged between 0.5 and 1.5. Two values were above 1.5, possibly indicating outliers. One of these two items was close to 2.0 and it was flagged for revision because it also had a separation index below the threshold of 0.15.

Table 11 shows that a total of 5 out of 75 items were flagged for revision, either because they were too difficult, too easy, or because they had a separation index below 0.15: 1 IM, 3 AM, and

1 Superior items. Poor separation values often coincided with poor difficulty values. Additionally, 5 items were flagged for inspection: 2 IL, 1 IM, 1 AM, and 1 Superior items.

A total of 20 items were inspected and a total of 12 were revised during the French LPT 01 revision process, while a total of 19 items were inspected and a total of 11 items were revised during the French LPT 02 revision process. Table 12 shows the item characteristics for German LPT 02.

Table 12
Item Characteristics German LPT 02

Item	Number of Cases	Measure	SEM	Infit	Outfit	Separation Index	Comment
A1.1.1	396	-3.47	.19	1.06	1.11	.14	Separation index below threshold. Revise.
A1.1.2	396	-.70	.11	1.05	1.05	.34	
A1.1.3	396	-.99	.11	1.03	1.04	.35	
A1.2.1	396	-2.82	.15	.98	.89	.28	
A1.2.2	396	-2.60	.14	1.01	1.04	.26	
A1.2.3	396	-.01	.11	1.05	1.10	.35	
A1.3.1	396	-3.20	.17	1.08	1.32	.11	Separation index below threshold. Revise.
A1.3.2	396	-1.75	.12	1.12	1.24	.22	
A1.3.3	396	-1.98	.12	.94	.87	.38	
A1.4.1	396	-2.97	.16	.95	.90	.29	
A1.4.2	396	-3.11	.17	1.02	1.06	.21	
A1.4.3	396	-2.58	.14	.90	.86	.36	
A1.5.1	396	-3.33	.18	1.03	.94	.19	Separation index below threshold. Inspect.
A1.5.2	396	-4.34	.28	.97	.66	.19	Item too easy. Revise.
A1.5.3	396	-5.07	.38	.98	.50	.15	Item too easy. Revise.
A2.1.1	522	.74	.11	1.18	1.31	.26	
A2.1.2	522	-2.35	.13	.86	.70	.43	
A2.1.3	522	-1.11	.10	.81	.74	.56	
A2.2.1	522	-1.08A	.10	.80	.72	.52	
A2.2.2	522	-.73A	.10	.89	.85	.49	
A2.2.3	522	-1.87A	.11	.84	.75	.43	
A2.3.1	522	-.79	.10	1.01	.98	.40	
A2.3.2	522	-1.87	.11	1.00	1.18	.33	
A2.3.3	522	-.70	.10	.85	.79	.55	
A2.4.1	522	-1.51	.11	1.02	.98	.36	
A2.4.2	522	-3.86	.21	1.01	1.08	.15	Item too easy. Revise.
A2.4.3	522	-.82	.10	.76	.69	.61	
A2.5.1	522	-1.72	.11	.96	1.02	.38	
A2.5.2	522	1.29	.12	1.15	1.39	.25	Item too difficult. Revise.
A2.5.3	522	-1.29	.10	1.19	1.30	.22	
B1.1.1	870	-.43	.09	.95	.75	.48	
B1.1.2	871	.25	.08	.99	.96	.46	

B1.1.3	868	.31	.08	.89	.82	.54	
B1.2.1	870	.77	.08	.98	1.04	.46	
B1.2.2	869	.91	.08	.94	.99	.50	
B1.2.3	868	-.56	.10	.98	.85	.44	
B1.3.1	869	-.59	.10	.84	.68	.54	
B1.3.2	869	-.39	.09	.85	.68	.55	
B1.3.3	868	-.30	.09	.88	.79	.52	
B1.4.1	870	-.28A	.09	.94	.81	.37	
B1.4.2	870	.71A	.08	.95	.95	.51	
B1.4.3	864	1.24A	.08	.86	.88	.56	
B1.5.1	872	.66	.08	.95	.91	.50	
B1.5.2	870	.32	.08	.87	.90	.55	
B1.5.3	871	.10	.08	.89	.82	.53	
B2.1.1	652	.85	.09	1.11	1.16	.25	
B2.1.2	652	.76	.09	.80	.71	.58	
B2.1.3	652	2.30	.09	.96	1.01	.39	
B2.2.1	652	2.52	.09	1.15	1.43	.15	Separation index below threshold. Inspect.
B2.2.2	652	2.81	.09	1.24	1.42	.07	Separation index below threshold. Revise.
B2.2.3	651	2.97	.10	.99	1.09	.32	Item too difficult. Revise.
B2.3.1	652	3.20	.10	1.07	1.26	.20	Item too difficult. Revise.
B2.3.2	650	.65	.09	.99	.99	.38	
B2.3.3	652	.36	.10	1.10	1.16	.24	
B2.4.1	652	1.30A	.09	1.02	1.03	.36	
B2.4.2	651	1.34A	.09	1.10	1.09	.35	
B2.4.3	651	.30A	.10	1.01	.94	.42	
B2.5.1	652	1.99	.09	.98	.95	.40	
B2.5.2	652	.39	.10	1.01	.98	.34	
B2.5.3	652	1.27	.09	.88	.85	.51	
C1.1.1	608	.69	.09	1.01	.98	.36	
C1.1.2	608	1.14	.09	1.01	1.02	.36	
C1.1.3	604	3.08	.10	.93	.95	.38	
C1.2.1	608	2.37	.09	1.17	1.27	.18	Separation index below threshold. Inspect.
C1.2.2	607	3.15	.10	1.04	1.31	.24	
C1.2.3	608	1.09	.09	1.00	1.01	.37	
C1.3.1	609	2.40	.09	1.18	1.38	.14	Separation index below threshold. Revise.
C1.3.2	608	2.34	.09	1.00	1.13	.34	
C1.3.3	609	2.21	.09	.96	.99	.40	
C1.4.1	609	.28	.10	.99	.98	.35	Item may be too easy. Inspect.
C1.4.2	610	.96	.09	1.06	1.05	.31	
C1.4.3	607	2.21	.09	1.15	1.26	.20	
C1.5.1	608	2.08	.09	1.13	1.18	.23	
C1.5.2	607	1.59	.09	.98	.98	.40	
C1.5.3	600	2.53	.09	1.03	1.09	.31	

Table 12 shows the item characteristics for German LPT 02. It shows that the overall item difficulty increased with the sublevels tested as expected. The precision of the item difficulty parameter was high, as suggested by the SEM, varying from .11 to .28 at the IL level except for one item, which was flagged for revision because it was much too easy (A1.5.3); from .10 to .21 at the IM level, from .08 to .10 at the AL level, from .09 to .10 at the AM level, and from .09 to .10 at the Superior level. All infit and outfit values were between 0.5 and 1.5 and many of them were close to 1.0, indicating that the items fit the model well.

Table 12 shows that a total of 10 out of 75 items were flagged for revision, either because they were too difficult, too easy, or because they had a separation index below 0.15: 4 IL, 2 IM, 3 AM, and 1 Superior items. Poor separation values often coincided with poor difficulty values. Additionally, 4 items were flagged for inspection: 1 IL, 1 AM, and 2 Superior items.

A total of 15 items were inspected and a total of 11 were revised during the German LPT 01 revision process. The item analysis for German LPT 02 was completed for this report and the revision process will begin within the next few weeks. Table 13 shows the item characteristics for Spanish LPT 02.

Table 13
Item Characteristics Spanish LPT 02

Item	Number of Cases	Measure	SEM	Infit	Outfit	Separation Index	Comment
A1.1.1	487	-1.79	.11	.97	1.00	.35	
A1.1.2	487	-2.32	.12	.97	1.08	.30	
A1.1.3	487	1.55	.15	.93	.88	.34	Item too difficult. Revise.
A1.2.1	487	-2.81	.14	.85	.62	.46	
A1.2.2	487	-.47	.10	.85	.84	.52	
A1.2.3	487	-.08	.10	.99	.99	.37	
A1.3.1	487	-1.45	.10	.85	.77	.53	
A1.3.2	487	.89	.12	.92	.85	.40	
A1.3.3	487	-2.29	.12	.90	.76	.43	
A1.4.1	487	-1.77	.11	1.01	1.01	.31	
A1.4.2	487	-.09	.10	.88	.84	.49	
A1.4.3	487	-1.11	.10	1.03	1.02	.33	
A1.5.1	487	-.96	.10	1.03	1.04	.32	
A1.5.2	487	-.92	.10	.91	.88	.47	
A1.5.3	487	-.06	.10	1.10	1.10	.24	
A2.1.1	930	-.01	.07	1.15	1.20	.26	
A2.1.2	930	-1.27	.08	.89	.82	.48	
A2.1.3	930	-1.48	.08	1.14	1.23	.21	Item may be too easy. Inspect.
A2.2.1	930	-1.13	.08	1.07	1.07	.31	
A2.2.2	930	-.87	.07	.85	.79	.53	
A2.2.3	930	-.08	.07	1.00	.99	.41	
A2.3.1	930	-.36	.07	.98	.99	.42	

A2.3.2	930	1.37	.09	1.23	1.78	.08	Separation index below threshold. Revise.
A2.3.3	930	-.86	.07	1.00	.97	.40	
A2.4.1	930	-.82A	.07	.78	.71	.57	
A2.4.2	930	-1.07A	.08	1.19	1.22	.34	
A2.4.3	930	.63A	.08	1.03	1.18	.40	
A2.5.1	930	.54	.08	1.04	1.14	.34	
A2.5.2	930	-.34	.07	1.09	1.09	.32	
A2.5.3	930	.01	.07	.85	.86	.54	
B1.1.1	530	-2.34	.16	.88	.58	.36	Item too easy. Revise.
B1.1.2	530	-1.61	.13	.99	1.12	.25	Item may be too easy. Inspect.
B1.1.3	530	.90	.10	.95	.98	.42	
B1.2.1	530	.51A	.09	1.09	1.11	.26	
B1.2.2	530	1.03A	.10	1.03	1.07	.29	
B1.2.3	530	.53A	.09	.90	.89	.47	
B1.3.1	530	.43	.09	.85	.81	.54	
B1.3.2	530	-.03	.09	1.14	1.18	.20	
B1.3.3	530	.50	.09	.89	.88	.49	
B1.4.1	530	-1.05	.11	.99	.99	.31	
B1.4.2	530	.06	.09	1.10	1.13	.25	
B1.4.3	530	1.76	.11	1.36	1.76	-.10	Separation index below threshold. Revise.
B1.5.1	530	-.65	.10	1.10	1.22	.19	Separation index below threshold. Inspect.
B1.5.2	530	.63	.09	1.07	1.10	.28	
B1.5.3	530	1.40	.10	.95	.95	.40	
B2.1.1	108	-.08	.22	.92	.88	.44	
B2.1.2	108	-1.29	.30	1.02	.80	.27	Item too easy. Revise.
B2.1.3	108	.88	.21	1.00	.99	.40	
B2.2.1	108	.64A	.21	.89	.83	.54	
B2.2.2	108	.96A	.21	.85	.82	.55	
B2.2.3	108	1.64A	.22	.95	.89	.56	
B2.3.1	108	-1.70	.34	.93	.65	.33	Item too easy. Revise.
B2.3.2	108	2.08	.24	.82	.71	.56	
B2.3.3	108	-.52	.24	.93	.75	.43	Item may be too easy. Inspect.
B2.4.1	108	1.57	.22	1.26	1.28	.15	Separation index below threshold. Inspect.
B2.4.2	108	1.38	.22	1.11	1.12	.30	
B2.4.3	108	2.70	.27	1.17	1.39	.14	Separation index below threshold. Revise.
B2.5.1	108	-.97	.27	.95	.83	.35	Item may be too easy. Inspect.
B2.5.2	108	.71	.21	.92	.88	.48	
B2.5.3	108	1.86	.23	.88	.83	.51	
C1.1.1	23	.51	.51	1.05	1.42	.24	
C1.1.2	23	2.03	.45	1.08	1.13	.27	
C1.1.3	23	-.94	.76	.92	.54	.38	Item too easy. Revise.
C1.2.1	23	-1.71	1.04	1.14	4.56	-.30	Separation index below threshold. Revise.
C1.2.2	23	1.21	.46	1.52	1.57	-.17	Separation index below threshold. Revise.
C1.2.3	23	-.94	.76	1.12	1.16	.09	Separation index below threshold. Revise.
C1.3.1	23	1.63	.45	1.09	1.10	.28	
C1.3.2	23	.99	.47	1.11	1.02	.29	
C1.3.3	23	1.63	.45	.92	.89	.47	
C1.4.1	23	.51	.51	.94	1.30	.35	

C1.4.2	23	-.45	.65	.81	.48	.53	Item too easy. Revise.
C1.4.3	23	.24	.54	1.05	.90	.32	
C1.5.1	23	2.24	.46	1.03	1.10	.31	
C1.5.2	23	.24	.54	1.07	1.21	.23	
C1.5.3	23	1.21	.46	1.08	1.10	.29	

Table 13 shows the item characteristics for Spanish LPT 02. It shows that the overall item difficulty increased with the sublevels tested as expected. The precision of the item difficulty parameter was high, as suggested by the SEM, varying from .10 to .15 at the IL level, from .07 to .08 at the IM level, from .09 to .16 at the AL level, from .21 to .34 at the AM level, and from .46 to .76 at the Superior level except one item, which has a SEM of 1.04. The high SEM values at the Superior level are due to the very small number of examinees who had taken Superior items. These SEM values indicate that item difficulty measures and separation indices may not be very reliable. All but one infit values were between 0.5 and 1.5 and many of them were close to 1.0, indicating that the items fit the model well. Three outfit values were between 1.5 and 2.0. All three of them were flagged for revision (red) because they also had separation indices below the threshold of 0.15. Item C1.2.1., which had a SEM above 1 had an outfit value of 4.56, clearly indicating that something was wrong with the item.

Table 13 shows that a total of 11 out of 75 items were flagged for revision, either because they were too difficult, too easy, or because they had a separation index below 0.15: 1 IL, 1 IM, 2 AL, 3 AM, and 4 Superior items. Poor separation values often coincided with poor difficulty values. Additionally, 6 items were flagged for inspection: 1 IM, 2 AL, 3 AM, and 1 Superior items.

A total of 15 items were inspected and a total of 12 were revised during the Spanish LPT 01 revision process, while a total of 17 items were inspected and a total of 13 items were revised during the Spanish LPT 02 revision process. This included all the items flagged for revision in Table 13 and two additional ones. Both French LPT 02 and Spanish LPT 02 were recently revised and are, at present, undergoing UAT to replace the present tests within a few weeks. German LPT 02 is slated for revision.

14 Reliability Information

To measure the internal consistency of the five sublevels, Cronbach's Alpha was computed for all examinees who took the complete test, i.e. who completed all five sublevels (Version H). Cronbach's Alpha provides an overall reliability estimate and is considered to be a measure of scale reliability. A value above 0.8 suggests that the items have relatively high internal consistency. Table 14 shows Cronbach's Alpha for the examinees who took all five sublevels.

Table 14
Scale Reliability of the French, German, and Spanish LPTs

Language	*N*	Cronbach's Alpha
French	91	0.852*
German	19	0.908*
Spanish	454	0.889*

* $p < 0.5$

Table 14 shows that Cronbach's Alpha is above 0.8 for all languages, indicating relatively high internal consistency of the items.

The Rasch person separation reliability was calculated for the whole test as another reliability measure. As suggested by AREA/APA/NCME (2014: 46), both, the overall and conditional standard errors of measurement (SEM) are considered to be central indicators of test reliability. The Rasch person separation reliability is considered to be equivalent to Cronbach's alpha. The Rasch person separation reliability, however, is sample independent and tends to underestimate the true reliability, whereas classical measures such as Cronbach's alpha tend to overestimate the true reliability. Note that the following analysis was based on French LPT 01, German LPT 01, and Spanish LPT 01. The number of examinees of this analysis was different from the number of examinees in the final data reports (Appendix 7), because the study took place at a different date. Table 15 presents overall Rasch separation reliability estimates as well as the conditional SEMs for the four two-sublevel tests.

Table 15
Reliability Estimates of the ACTFL Listening Proficiency Test (LPT)

	N	Overall SEM	Rasch Separation Reliability	Conditional SEM			
				IL/IM	IM/AL	AL/AM	AM/S
French	816	.42	.81	.45 (*N* = 363)	.43 (*N* = 270)	.43 (*N* = 241)	.42 (*N* = 93)
German	239	.45	.84	.46 (*N* = 109)	.47 (*N* = 75)	.45 (*N* = 47)	n.a.*
Spanish	1884	.43	.83	.45 (*N* = 735)	.46 (*N* = 419)	.45 (*N* = 816)	.43 (*N* = 276)

* Not enough cases to calculate a meaningful SEM or meaningful difficulty estimates.

Table 15 shows that the overall Rasch person separation reliability is very high for all languages. The large majority of examinees took tests consisting of 30 items. The smallest SEM value possible for a test with 30 items is 0.37. The observed overall SEMs are only marginally higher than this, indicating a high degree of reliability for the number of items used. The conditional SEMs are equally low. The measures reported in this table, therefore, provide additional evidence that the LPT has a high degree of reliability.

This conclusion is corroborated by the overall Rasch item fit statistics in Table 16 (see Section 13 for item fit statistics for individual items and see Appendix 7 for the fit statistics reported in Table 16).

Table 16
Overall Rasch Fit Statistics

	N	**Rasch Item Infit (MNSQ)**	**Rasch Item Outfit (MNSQ)**
French LPT 01	1,127	1.00	1.03
French LPT 02	666	0.98	1.02
German LPT 01	342	1.00	1.06
German LPT 02	661	0.99	0.99
Spanish LPT 01	1,769	1.01	1.07
Spanish LPT 02	1,185	1.00	1.05

Table 16 shows that the items generally produce exactly the same amount of infit variance that would be expected from the Rasch model. Outfit values are equally close to the ideal variance range. The overall Rasch fit statistics, thus, add another piece of evidence to support the conclusion that the measurement functions as desired.

15 Evidence for the Equivalence of Forms of the Test

There are several measures in place to ensure equivalence of test forms: the training and monitoring of item writers and reviewers; the use of anchor items; and the revision of test forms on the basis of IRT analysis.

Item writers and reviewers are rigorously trained and monitored throughout the passage and item writing process (see *Item Development Process* in Section 3) and they are provided with a very detailed Item Writing Manual and Item Checklists (see Appendix 9 – Item Writing Manual and Appendix 10 – Item Checklists). The same item writers and/or reviewers are commonly involved in several test forms. Because the passages and items are reviewed and revised at least twice and because there are at least three experienced item writers and reviewers involved in

every single test form, there is a precise and deeply shared understanding of what the ACTFL levels and sublevels involve.

Three anchor passages and nine anchor items of one form are used for any subsequent form, i.e. three anchor items at IM, three at AL, and three at AM. These anchor items are carefully selected on the basis of the IRT analysis and exhibit the best difficulty measures and separation indices of a particular form. By means of common item equating using the WINSTEPS software, the difficulty of new test items is determined with high precision.

IRT analyses are completed for all forms after 300-400 test administrations. Items with irregular values are inspected and revised, if necessary (see Section 13). This is a mandatory part of the item development cycle. Revised forms become part of the form pool and will be inspected and revised on the basis of other IRT analyses further down the road. These revisions ensure even greater form equivalency.

Figures 1, 2, and 3 show logit boxplots of the two French, two German, and two Spanish forms available at present. Note that this analysis is based on forms that have not been revised. (All but German LPT 02 have been revised as of this writing and will replace the current versions within a few weeks.)

Figure 1
Logit Distribution of French LPT 01 and LPT 02

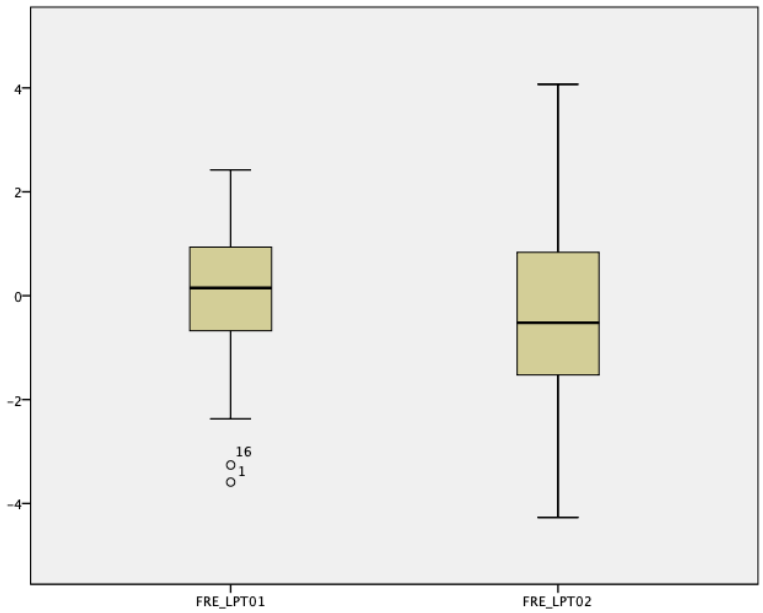


Figure 2
Logit Distribution of German LPT 01 and LPT 02

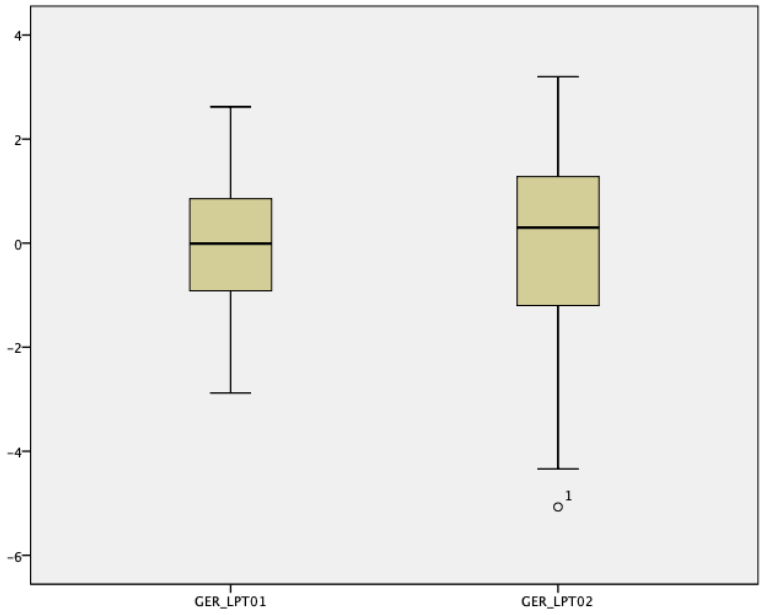
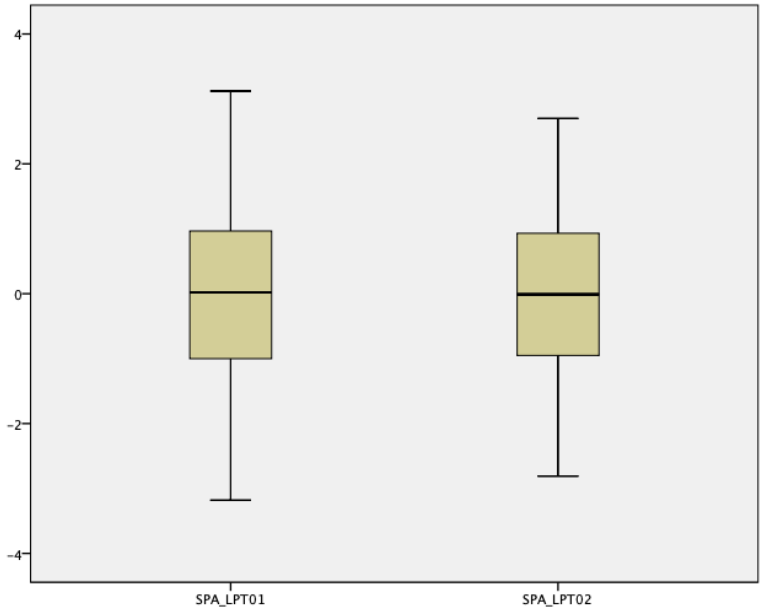


Figure 3
Logit Distribution of Spanish LPT 01 and LPT 02



Figures 1, 2, and 3 show very similar distributions for the two forms of each language. For French and German, the medians were similar as were the interquartile ranges (IRQ) (boxes). Full ranges (whiskers) were a little more pronounced for form 2. For Spanish, medians, IRQs and full ranges were almost identical. This provides evidence of form equivalence even before the first mandatory revision, i.e. on the basis of the quality of the item development process alone. After the revisions completed on the basis of the full analyses of all forms (see Appendix 7 – French, German, and Spanish LPT Data Reports for all French, German, and Spanish forms), the form equivalence will be even greater.

Table 17 shows the number of test administrations, logit medians, means, standard errors of the mean (SEM), and standard deviations of all seven test forms.

Table 17
Descriptive Statistics of two French, two German, and two Spanish LPT Forms

	N	Median	Mean	SEM	SD
French LPT 01	75	0.15	0.00	0.15	1.31
French LPT 02	75	-0.52	-0.38	0.20	1.71
German LPT 01	75	-0.01	0.00	0.14	1.23
German LPT 02	75	0.30	-0.04	0.23	1.97
Spanish LPT 01	75	0.02	0.00	0.15	1.31
Spanish LPT 02	75	-0.01	0.01	0.15	1.26

Table 17 shows that the logit medians, means, and standard deviations of the two French, two German, and three Spanish forms are very similar to each other. The means for French LPT 01 and 02, e.g., are 0.38 logits apart and the difference in standard deviations is 0.4. German and Spanish means vary by 0.04 and 0.01 logits. The standard deviations for German vary by 0.74, while the standard deviations for Spanish vary by only 0.05. The statistics in Table 17, thus, support the claim that test forms are identical in Spanish, very similar in German and slightly less so in French. All tests except German LPT 02 have been revised and are currently undergoing UAT. When the current test versions will have been replaced, it is expected that the test form equivalence for French will be strong as well.

16 Scorer Reliability for Essay Items

Not applicable

17 Errors of Classification Percentage for the Minimum Score for Granting College Credit (Cut-Score)

Table 18 shows the logits and their respective standard error of measurement (SEM) of all cut scores distinguishing between ACTFL LPT sublevels (see Appendix 8 for logits and SEMs for all scores from 1 to 75 for French LPT 01, German LPT 01, and Spanish LPT 01). Cut-score logits and SEMs are calculated on the assumption of an examinee having responded to all 75 items of a complete test.

Table 18
Cut-Score Logits and SEMs for All ACTFL Levels by Language

		French LPT 01		German LPT 01		Spanish LPT 01	
ACTFL	**Cut-score**	**Logit**	**SEM**	**Logit**	**SEM**	**Logit**	**SEM**
NL	below 12						
NM	12	-2.18	.36	-2.15	.35	-2.19	.35
NH	15	-1.82	.33	-1.81	.32	-1.85	.32
IL	18	-1.50	.31	-1.52	.31	-1.55	.31
IM	24	-.96	.29	-1.00	.28	-1.02	.29
IH	37	.03	.27	-0.03	.27	-.03	.27
AL	48	.82	.27	.77	.28	.79	.28
AM	54	1.28	.29	1.26	.29	1.28	.30
AH	67	2.66	.39	2.70	.40	2.75	.41
S	69	3.00	.44	3.06	.45	3.12	.45

Table 18 shows that the SEM is low for all sublevel cut points and languages. The logits are similar for all three languages, indicating equivalence across languages. At ACTFL levels IL to AM, SEMs range from .27 to .31 for all three languages. At ACTFL levels AH and S, SEMs range from .39 to .45 for all three languages.

18 Evidence of Validity: Content-related

Each exam provides a representative sample of the construct by including a broad spectrum of topics, subtopics, genres, and rhetorical organization (text type). The LPT is commonly taken as a two-sublevel test and consists of ten passages, five at each level. The ten passages are chosen to provide a representative statement of the language proficiency of the examinee. In Section 8, three examples of different two-level tests were presented to show how the passages reflect the *ACTFL Proficiency Guidelines 2012 – Listening*, and how the test ensures selecting a diverse and representative sample of the topics, subtopics, genres, and rhetorical organization of passages listeners need to be able to understand to be rated at a particular proficiency for each level. These examples showed that the tasks in any single exam cover a broad spectrum of top-

ics, subtopics, genres and rhetorical organization and provide a solid and representative statement of the listening proficiency of examinees.

19 Evidence of Validity: Criterion-related

The ACTFL LPT was externally validated by a side-by-side study with NATO's Benchmark Advisory Test – Listening (BAT-L) (see Appendix 6 – Technical Report). The present section describes the analyses that were carried out to determine the **internal validity** of the ACTFL LPT as well as how insights about its **external validity** were gained.

Subjects and Instruments

The subjects were students of English at the University of Leipzig ranging from beginning to very advanced levels (Bärenfänger & Tschirner, 2013; Tschirner & Bärenfänger). Both the ACTFL LPT and NATO's BAT-L were administered to a total of 88 examinees. The BAT-L measures listening proficiency using the STANAG 6001 scale, which is derived from the ILR scale, the scale used by U.S. government agencies. The ILR scale was used as the basis for the ACTFL scale. Both scales continue to be commensurate, which means that there are precise correspondences between ACTFL and ILR levels.

To ensure a relatively even distribution of proficiency levels, an almost equal number of participants were selected from Beginning, Intermediate 1, Intermediate 2, and Advanced English courses. Also included in the sample were advanced students of English teacher education, American Studies, and Translation Studies to gain insights into the ACTFL Superior level. Since beginners in university language classes in Germany are relatively rare, the proportion of participants with beginning proficiency in English was smaller than that of participants with more advanced proficiency.

Research Design

Both, LPT and BAT-L were administered to the same group of students in a split test design. Half the participants took the LPT first; the other half took the BAT-L first. Participants took both tests internet-delivered under controlled proctored conditions in University of Leipzig computer labs. The tests were taken at different days to prevent participant fatigue. Lower proficiency students took LPT sublevels IL, IM, and AL and BAT-L levels 1 and 2. Mid-level proficiency students took LPT sublevels AL and AM and BAT-L levels 1 and 2. High-level proficiency students took LPT sublevels AL, AM, and S and BAT-L levels 2 and 3. Participants were given 75 minutes for the three-sublevel LPT and the BAT-L and 50 minutes for the two-sublevel LPT. Tests were computer-scored according to their internal scoring algorithms. For the three-sublevel LPT, the two highest levels that had at least sixty per cent of the items correct were scored to arrive at the final rating.

Statistical Analyses

To determine the *internal validity* of the LPT, two types of analyses were carried out. Within the framework of classical test theory, Cronbach's alpha was computed for each level of the test as a measure of overall reliability. In addition, information about the reliability of each individual item was collected by calculating item difficulty parameters and item discrimination parameters. Probabilistic test theory (Rasch dichotomous model) was used to provide a further perspective and to gain more fine-grained insights into the validity of the LPT.

To gain insights into the *external validity* of the ACTFL LPT, raw percentages of agreement between the LPT and BAT-L were cross-tabulated, and the following correlation values were computed: Raw percentage of agreement; Pearson's correlation; Spearman's *rho*; Kendall's *tau*; and Goodman and Kruskal's *gamma*.

Data Analysis

Table 19 displays all measures that were computed to establish the ACTFL LPT's *external validity*. It contains four parameters, which describe the relationship between the ACTFL LPT and the BAT-L. Two correlation and two agreement measures were computed. Both correlation parameters, Pearson's *r* and Spearman's *rho* show high interdependence between the two tests. As for the agreement measures, Kendall's *tau* is affected by bindings in the data and thus somewhat lower than Goodman-Kruskall's *gamma*. Both indicators support, however, the conclusion that there is high agreement between the ratings of both tests.

Table 19
Correlation and Agreement Measures Between Final Ratings of the ACTFL LPT and the BAT-L

N	**Pearson's *r***	**Spearman's *rho***	**Kendall's *tau***	**Goodman-Kruskall's *gamma***
88	.842*	.833*	.753*	.898*

*Correlations were significant at $p < 0.01$.

To confirm that the results of the LPT show the correct correspondences between ACTFL and ILR levels, the frequency distribution between the two sets of results was examined. Table 20 presents the frequency of agreement in final ratings between the LPT and the BAT-L.

Table 20
Frequency of Agreement in Final Ratings between the LPT and the BAT-L

		BAT-L Final Rating						
		0	0+	1	1+	2	2+	3
ACTFL LPT Final Rating	0	1 (1.0)						
	IL		2 (.40)	3 (.60)				
	IM			8 (.57)	3 (.21)	3 (.21)		
	AL			3 (.09)	8 (.23)	23 (.66)		
	AM			1 (.14)		1 (.14)	2 (.29)	3 (.43)
	S					4 (.15)	6 (.23)	16 (.62)

*Note: The proportion of agreement is indicated in parentheses.

Table 20 shows that the following correspondences between the results of the two tests had the greatest proportion of agreement: IL and ILR 1 (60%); IM and ILR 1 (57%); AL and ILR 2 (66%); S and ILR 3 (62%). This represents the relationship between ACTFL and ILR well. The established correspondences between ACTFL and ILR are as follows: IL corresponds to ILR 1; IM (rarely IH) corresponds to ILR 1+; AL corresponds to ILR 2; AM (rarely AH) corresponds to ILR 2+; and baseline Superior corresponds to ILR 3.

The finding that IL agrees with 0+ (40%) and ILR 1 (67%), i.e. the lower ILR 1 ranges, and that IM agrees with ILR 1 (57%) and ILR 1+ (21%), i.e., the higher level 1 ranges, is consistent with the relationship between ACTFL and ILR as established above as. Similarly, the finding that AL corresponds to ILR 1+ (23%) and ILR 2 (66%), i.e., the lower ILR 2 ranges, and AM corresponds to ILR 2 (14%) and ILR 2+ (29%) and even to ILR 3 (43%), i.e., the higher level 2 ranges, is also consistent with the established relationship between ACTFL and ILR. Superior, finally, clearly corresponds to ILR 3 (62%). The results of this study, therefore, provide external validity evidence, i.e. criterion-related validity evidence.

20 Evidence of Validity: Construct-related

There are two pieces of evidence to support the construct validity of the LPT: The results of a standard-setting workshop; and the Rasch model fit.

Standard-setting Workshop

The first piece of evidence comes from a two-day standard-setting workshop, which was conducted with the German LPT 01 in July 2015. Eight experts with a college degree in German as a

Foreign Language and with broad experience teaching and testing German as a Foreign Language participated in the study (one male and seven female). Employing the modified Angoff method (Impara & Blake, 1997), the experts were asked to judge each of the 75 items of one form of the German LPT whether a borderline candidate at a specific competence level would be able to answer test items at his or her competence level correctly.

The workshop consisted of three phases: familiarization, calibration, and standard-setting. In the familiarization phase, the experts ordered relevant competence descriptors in small groups and discussed their results. In addition, they discussed the salient features of each proficiency level. The overall aim of the familiarization phase, which lasted 90 minutes, was to create a shared understanding of the proficiency scale and the test construct.

In the calibration phase, participants applied their understanding of the listening proficiency construct individually to ten listening tests of German as a Foreign Language with calibrated difficulties (the tests included tests from the Goethe Institute, The European Language Certificates/telc, and Test-DaF). In the ensuing discussion, participants were asked to explain their judgments. There was high agreement among the participants with respect to the proficiency levels of the tests rated. The calibration phase lasted 90 minutes.

The standard-setting phase lasted 240 minutes. Participants were first asked to listen to an LPT passage and read its items. Then they were asked to judge whether a borderline candidate would be able to answer each of the three items correctly. Participants were also asked to indicate on a four-point Likert scale how confident they were of their rating. At the bottom of their rater sheets, they were able to comment on the passage, the items, and the rating process. The listening passages and items were ordered in two different orders: one set started with the easiest passages and continued to the more difficult ones, and the other set started with the most difficult passages and continued to the easier ones. This was intended to mitigate ordering effects. After the participants had judged all 75 test items, they were asked to comment on the rating process on a separate sheet.

Table 21 presents the results of the standard setting for each individual item. The first row of a group provides the item ID, the second row the number of participants, the third row the mean participant agreement on whether a borderline candidate would answer the item correctly ("yes" was coded "1", "no" was coded "0"), and the fourth row the standard deviation of the agreement measure.

Table 21
Results of the Standard-setting Workshop of the German LPT 01

	A1.1.1	**A1.1.2**	**A1.1.3**	**A1.2.1**	**A1.2.2**	**A1.2.3**	**A1.3.1**	**A1.3.2**	**A1.3.3**
N	8	8	8	8	8	8	8	8	8
Agreement	0.63	0.75	0.50	0.75	1.00	0.63	1.00	0.88	0.38
SD	0.52	0.46	0.53	0.46	0.00	0.52	0.00	0.35	0.52
	A1.4.1	**A1.4.2**	**A1.4.3**	**A1.5.1**	**A1.5.2**	**A2.1.3**	**A2.1.1**	**A2.1.2**	**A2.1.3**

N	8	8	8	8	8	8	8	8	8
Agreement	1.00	1.00	0.63	1.00	0.38	0.88	0.75	0.75	1.00
SD	0.00	0.00	0.52	0.00	0.52	0.35	0.46	0.46	0.00
	A2.2.1	**A2.2.2**	**A2.2.3**	**A2.3.1**	**A2.3.2**	**A2.3.3**	**A2.4.1**	**A2.4.2**	**A2.4.3**
N	8	8	8	8	8	8	8	8	8
Agreement	0.75	1.00	0.38	0.75	0.75	1.00	0.88	1.00	0.50
SD	0.46	0.00	0.52	0.46	0.46	0.00	0.35	0.00	0.53
	A2.5.1	**A2.5.2**	**A2.5.3**	**B1.1.1**	**B1.1.2**	**B1.1.3**	**B1.2.1**	**B1.2.2**	**B1.2.3**
N	8	8	8	8	8	8	8	8	8
Agreement	0.75	1.00	0.88	0.88	0.38	0.50	0.75	0.13	0.88
SD	0.46	0.00	0.35	0.35	0.52	0.53	0.46	0.35	0.35
	B1.3.1	**B1.3.2**	**B1.3.3**	**B1.4.1**	**B1.4.2**	**B1.4.3**	**B1.5.1**	**B1.5.2**	**B1.5.3**
N	8	8	8	8	8	8	8	8	8
Agreement	0.88	0.75	0.63	0.88	1.00	0.63	0.38	1.00	0.50
SD	0.35	0.46	0.52	0.35	0.00	0.52	0.52	0.00	0.53
	B2.1.1	**B2.1.2**	**B2.1.3**	**B2.2.1**	**B2.2.2**	**B2.2.3**	**B2.3.1**	**B2.3.2**	**B2.3.3**
N	8	8	8	8	8	8	8	8	8
Agreement	1.00	0.88	0.50	0.75	0.75	0.88	0.50	0.50	0.25
SD	0.00	0.35	0.53	0.46	0.46	0.35	0.53	0.53	0.46
	B2.4.1	**B2.4.2**	**B2.4.3**	**B2.5.1**	**B2.5.2**	**B2.5.3**	**C1.1.1**	**C1.1.2**	**C1.1.3**
N	8	8	8	8	8	8	8	8	8
Agreement	0.88	0.75	0.88	1.00	0.75	0.75	1.00	0.63	0.75
SD	0.35	0.46	0.35	0.00	0.46	0.46	0.00	0.52	0.46
	C1.2.1	**C1.2.2**	**C1.2.3**	**C1.3.1**	**C1.3.2**	**C1.3.3**	**C1.4.1**	**C1.4.2**	**C1.4.3**
N	8	8	8	8	8	8	8	8	8
Agreement	0.88	1.00	0.38	0.63	0.75	0.75	1.00	0.75	0.75
SD	0.35	0.00	0.52	0.52	0.46	0.46	0.00	0.46	0.46
	C1.5.1	**C1.5.2**	**C1.5.3**						
N	8	8	8						
Agreement	0.75	0.88	1.00						
SD	0.46	0.35	0.00						

Rater agreement of 0.5 and higher indicates that the majority of raters believed that the item matches the test construct of a particular sublevel. As Table 21 shows, there were 8 out of 72 cases, where the raters judged an item too difficult for the targeted proficiency level; in all other cases, raters agreed with the level the item was supposed to target. This finding provides evidence of the alignment of the test with the construct matrix and proficiency scale.

Rasch Model Fit

The second piece of evidence of the construct validity of the LPT comes from Rasch measurement. Rasch statistics impose a theoretical model – in this case the Rasch model for dichotomous items – on empirical data. When the observed data fit the theoretical model, this may be

interpreted as an indication of the validity of the model, i.e. of construct validity. Rasch person infit and outfit values for each test form was provided in Table 16 in Section 14. For ease of reference, it is repeated in Table 22. A value of 1.0 implies a perfect fit, while values between 0.5 and 1.5 are considered to be an acceptable fit.

Table 22
Rasch Person Infit and Outfit Values

	N	**Rasch Item Infit (MNSQ)**	**Rasch Item Outfit (MNSQ)**
French LPT 01	1,127	1.00	1.03
French LPT 02	666	0.98	1.02
German LPT 01	342	1.00	1.06
German LPT 02	661	0.99	0.99
Spanish LPT 01	1,769	1.01	1.07
Spanish LPT 02	1,185	1.00	1.05

As Table 22 shows, the data fit the model impressively well. All infit values are within 0.02 MNSQ of a perfect fit of 1.0. All outfit values are within 0.07 MNSQ of a perfect fit. All test forms, therefore, are highly predictive of examinees' performance. This provides strong evidence of the construct validity of the test.

21 Possible Test Bias of the Total Test Score

Two main aspects for possible test bias are gender-based and culture-based bias. The item writing manual and the two check lists require writers and reviewers to keep these sources of bias in mind when writing and reviewing passages and items. Topics and items are developed to have equal appeal to both genders and they are developed and reviewed equally by female and male authors to avoid gender-based bias.

To avoid discrimination of certain cultures, causing culture-based test bias, emotionally charged topics such as sexuality, religion, war and violence as well as topics that are culture-specific are avoided, as is the use of inappropriate language.

Because LTI does not request nor collect personal information from examinees for privacy reasons (see Section 25), it is not possible to calculate differential item functioning (DIF) statistics. The steps outlined in Section 11, therefore, have been put in place to avoid including biased test items before operational testing.

22 Evidence that Time Limits are Appropriate and That the Exam is not Unduly Speeded

To determine if time limits are appropriate and the exam is not unduly speeded, the time it took examinees to finish the test was examined. The maximum amount of time provided to examinees for the standard two-sublevel test is 50 minutes. Table 23 shows the minimum, maximum, mean, standard error of the mean (SEM), and standard deviation (SD) of the time in minutes it took the examinees to take the test per language. In addition, Table 23 shows the percentage of examinees who used the full 50 minutes.

Table 23
Number of Test-Takers by Language, Minimum, Maximum, Mean, and Standard Deviation of Time it Took to Complete the Test, and Percentage of Test-takers who took the full 50 minutes

Language	*N*	Minimum	Maximum	Mean	SEM	SD	50 min
French	1781	20	50	29.94	0.12	5.03	0.5%
German	614	16	46	25.73	0.24	5.82	0.0%
Spanish	4291	14	50	27.92	0.09	5.81	0.2%

Table 23 shows that very few examinees take the full amount of 50 minutes. Less than 99% of the examinees in French and Spanish and no one in German took the full 50 minutes. This may be taken as evidence that the time limits are appropriate and that the test is not unduly speeded. The average time it took examinees to take the test was 30 minutes in French, 26 minutes in German, and 28 minutes in Spanish.

23 Provisions for Standardizing Administration of the Examination

This section summarizes the provisions for standardizing the administration of the examination (see Appendix 2 – Assessment Use Argument and Appendix 11 – Examinee Handbook). Impartial treatment of examinees during all aspects of the administration of the LPT from registering for the assessment to taking the assessment is ensured by making sure the following regulations are adhered to.

- Individuals have equal access to information about LPT content and procedures.
- Individuals have equal access to the LPT, in terms of cost, location, and familiarity with conditions and equipment.
- Individuals have equal opportunity to demonstrate the ability to be assessed.

Examinees may access information about the test and download the LPT Familiarization Manual and the Examinee Handbook from the official homepage of Language Testing International (LTI), the ACTFL Testing Office.

The LPT is delivered over the Internet using the same test algorithm every single time and it is accessible to examinees in any part of the world where there is reliable Internet availability.

The LPT is a machine-scored test administered online. Official ACTFL LPT ratings are assigned to LPTs by LTI. Persons supervising the test are required to treat all examinees impartially following procedures described in the Examinee Handbook.

24 Provisions for Exam Security

Language Testing International (LTI), ACTFL's test administration office, has built test registration, scheduling, test management, and delivery test processing platforms that meet the high security standards for encrypting personal information and hosting tests on Amazon Web Services (AWS). Data is securely backed up in redundant locations in order to ensure 24/7 performance and data security.

At the completion of every test, answers are immediately streamed to a secured cloud datacenter, preventing the possibility of any response being stored. All servers are hardened for security and are also part of a high-availability cloud cluster. Cloud servers are managed and monitored by the data center, in conjunction with LTI, for performance and security events. Responses are backed up daily and the data is stored in a secure environment.

LTI's Client Site, a part of the aforementioned test management system, is a web-based portal that provides those who are registering for an ACTFL test with various options to register and monitor progress throughout the testing process, from pre-test to post-test administration. Access to LTI's Client Site is privilege-based and restricts modules' access to users based on their accounts' configuration. Users can: (1) request language tests; (2) view all of the tests that have been completed along with their results; (3) generate certificates of proficiency for relevant tests; and (4) update billing information.

All records are stored electronically in a secure environment. Examinees' names and assessment results are stored securely in LTI's database repository. All personally identifiable information is digitally encrypted to prevent unauthorized access. LTI's production servers are located in an SOC 2 compliant datacenter where access is secured using biometric access controls.

LTI intentionally uses only the minimum amount of data needed to take a test. LTI will not disclose any customer identifiable information (CII), such as customer name, home or email address, or phone number unless directed by the customer. LTI may use anonymous, aggregated information about its customers for internal research, or to update and/or maintain its systems. However, LTI does not sell, rent, or loan any CII to any third parties that are not authorized ser-

vice providers, or who are not clients with whom LTI has signed Confidentiality Agreements concerning the use of Customer Information. LTI's full privacy statement is located at: https://www.languagetesting.com/privacy.

25 Scaling and IRT Procedures

The IRT model used is the Rasch model for dichotomous items. All items are dichotomously scored as correct or incorrect. The Rasch model was selected because it allows person ability and item difficulty measures to be put on the same scale and because it works well with responses that consist of yes/no answers (correct/incorrect). The full model is used for scoring purposes in the ACTFL Listening and Reading Computer-Adaptive Test (L&Rcat). For the LPT (and RPT), the model is used for scaling new items on the old scale with the help of anchor items (see Section 15). See *Rasch Model Fit* in Section 20 for evidence that the items of the six LPT forms for French, German, and Spanish fit the Rasch model to a very high degree. To ensure that the results with the new items (new forms) have the same meaning and interpretation as the previous form, a total of nine anchor items are used. See Section 15 for evidence that the new items for each subsequent form fit both the IRT model and scale previously adopted and used.

26 Validity of Computer Administration

The ACTFL LPT was designed as a computer-administered test from the start. There are no paper-and-pencil versions. Examinees have to wait for 90 days before they can retake the test. LTI keeps track of which form an examinee took so that a different but equivalent form of the test can be used when they retake the test. Currently, there are four different but equivalent forms for French, two for German, and eight for Spanish. New forms are developed continually.

27 Cut-Score Information

Cut scores were determined empirically through a side-by-side study between the LPT and NATO's Benchmark Advisory Test – Listening (BAT-L) (See Section 19 and Appendix 6 – Technical Report). The BAT-L rates listening proficiency using the STANAG 6001 scale, which is derived from the ILR scale, the scale used by U.S. government agencies. The ILR scale was also used as the basis for the ACTFL scale resulting in precise correspondences between ACTFL and ILR levels.

The BAT-L uses a percentage system to convert scores to levels: 1-30% is considered a random effect (may, e.g., be achieved through guessing); 31-50% is considered emerging proficiency; 51-70% is called developing proficiency; and a score above 70% is considered as evidence of a proficiency level. The side-by-side study revealed that for the LPT, the percentages that aligned best with the results of the BAT-L were 40%, 60%, and 80%. Scores below 40%, i.e. below 12, were

found to be *random*, i.e. they indicated no evidence of any level; scores between 60% and 79%, i.e. between 18 and 23, were found to provide evidence of the examinee being at the lower of the two levels considered; and scores of 24 and above were found to provide evidence of the examinee having reached the higher of the two levels considered.

Because the LPT is a high stakes test, false positive classification decisions were considered to be relatively more serious than false negative classification errors. Therefore, cut scores were set at the upper end of the cut score range determined by the calibration study. (See Section 19 for more information on the study and the way cut-scores were determined).

These cut-scores were verified in a later study using another type of empirical data, the results of a standard-setting workshop relying on expert judgments (see Section 20). Table 24 displays the mean agreement of the expert judges across all items of the main proficiency sublevels of the test.

Table 24
Mean Rater Agreement on the Cut-Scores of the German LPT 01

	N	Cut-Score IL	Cut-Score IM	Cut-Score AL	Cut-Score AM	Cut-Score S
German	8	.76 (SD* = .32)	.81 (SD = .30)	.68 (SD = .39)	.73 (SD = .39)	.79 (SD = .34)

*SD = Standard Deviation

As Table 24 shows, the cut-scores as estimated in the standard-setting workshop were consistently in the range of 0.73 and 0.81 except for AL where the cut-score is slightly below .70. Because it seems safe to assume that an examinee has to answer at least 70% of the items of any proficiency sublevel correctly to be placed at this sublevel, these expert judgments provided further evidence of the reasonableness and appropriateness of the cut-scores recommended on the basis of the side-by-side study.

A third piece of evidence that the cut-scores are reasonable and appropriate comes from an analysis of the means of the two sublevels that are rated together. Because the algorithm simply counts the number correct of both sublevels, it is important to know which sublevel contributes most to a rating. While it may be safe to assume that it is not very relevant to distinguish between IL and IM items, which are relatively similar to begin with, when determining if an examinee is IL or IM, and that correct responses of both sublevels may simply be added together, this approach may need to be supported more substantially for test versions that combine two main levels such as version B, which combines IM and AL items, and version D, which combines AM and S items (see Section 1). Table 25 shows the number of test administrations, the mean score, the standard error of the mean (SEM), and the standard deviation of the two sublevels of all Spanish two-sublevel tests administered separately for the lower and the upper rating. The lower rating for Version A is IL and the upper rating is IM. For Version B, the lower rating is IM and the upper rating is AL. For version C, the lower rating is AL and the upper rating is AM. For Version D, the lower rating is AM and the upper rating is S.

Table 25
Number of Test Administrations, Mean Scores, SEM, and SD for all Spanish Two-Sublevel Tests Separate by Rating

		Lower Rating				Upper Rating			
Version		N	Mean	SEM	SD	N	Mean	SEM	SD
A	IL	382	11.32	0.08	1.56	50	13.20	0.17	1.18
	IM	382	8.55	0.92	1.80	50	12.08	0.18	1.29
B	IM	222	10.16	0.12	1.78	96	12.91	0.12	1.16
	AL	222	9.33	0.12	1.76	96	12.38	0.13	1.30
C	AL	358	11.71	0.07	1.40	90	13.27	0.10	0.96
	AM	358	8.46	0.09	1.62	90	11.78	0.11	1.06
D	AM	123	9.89	0.11	1.26	27	12.41	0.27	1.39
	S	123	9.56	0.12	1.33	27	12.37	0.21	1.08

performperformperformTable 25 shows that the mean scores are consistently higher for the lower level than for the higher level, and higher for the upper rating than for the lower rating. The latter is not surprising because one needs a higher score to be rated at the higher level. The former, however, is significant, because it means that examinees do, indeed, get a higher score at the lower level and a lower score at the higher level. Take Version A as an example. For the lower rating, i.e. IL, the average examinee score for IL items was 11.32 and for IM items, it was 8.55. This means that examinees generally had 11 out of 15 items correct at IL when their final rating was IL, while they had less than 9 out of 15 items correct at IM. This shows that IL items contributed more to an IL rating than IM items. Looking at the upper rating, i.e. IM, one sees that the mean IL score for all examinees rated IM was 13 out of 15. This means that examinees rated IM generally had almost all of the IL items correct. Table 25 shows that the items of Version C perform similarly and involve a clear distinction between AL and AM scores for examinees rated AL and AM. The situation is less clear for Versions B and D. Figures 4-8 show boxplots of all Spanish scores for these two versions.

Figure 4
Mean IM and AL Scores for Spanish Version B Examinees Rated IM

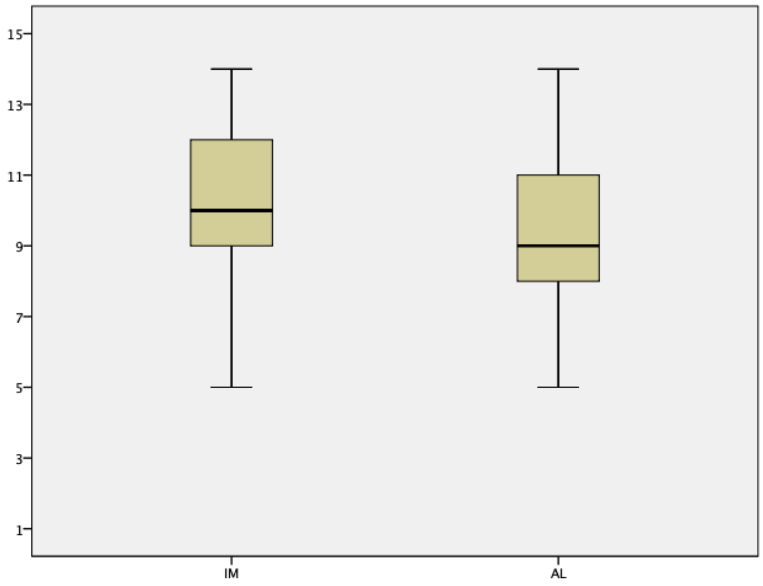
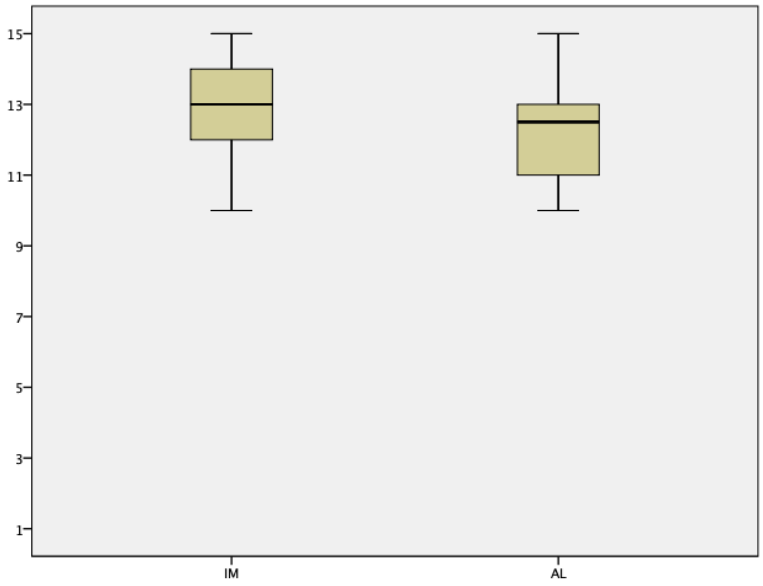


Figure 5
Mean IM and AL Scores for Spanish Version B Examinees Rated AL



Figures 4 and 5 show that the median and the interquartile range (IQR) are higher for IM items than for AL items for examinees who were rated IM and for examinees who were rated AL. More significantly, examinees who were rated AL had most of the IM items correct (median 13, IQR 12-14). A paired samples *t*-test (two-tailed) found that the difference between IM and AL mean scores was statistically significant supporting the conclusion that the cut-scores function as expected ($t = 4.37$, $p = 0.000$, $df = 317$).

Figure 6
Mean AM and S Scores for Spanish Version D Examinees Rated AM

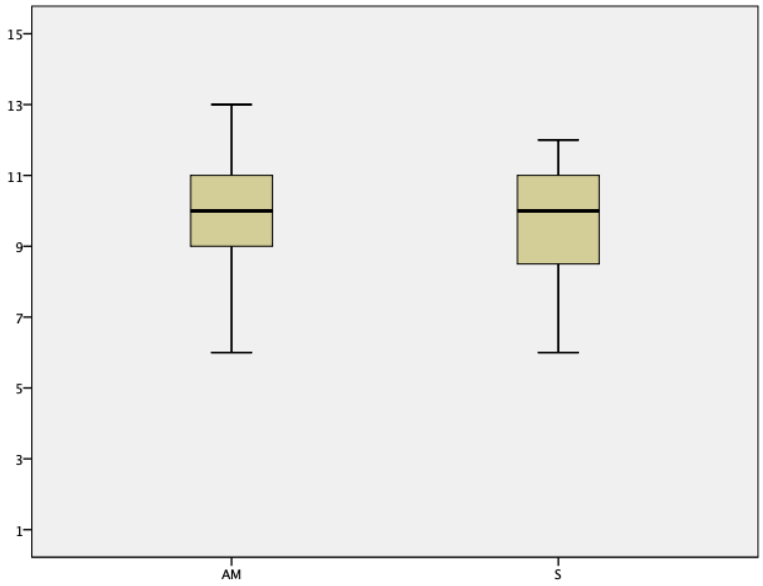
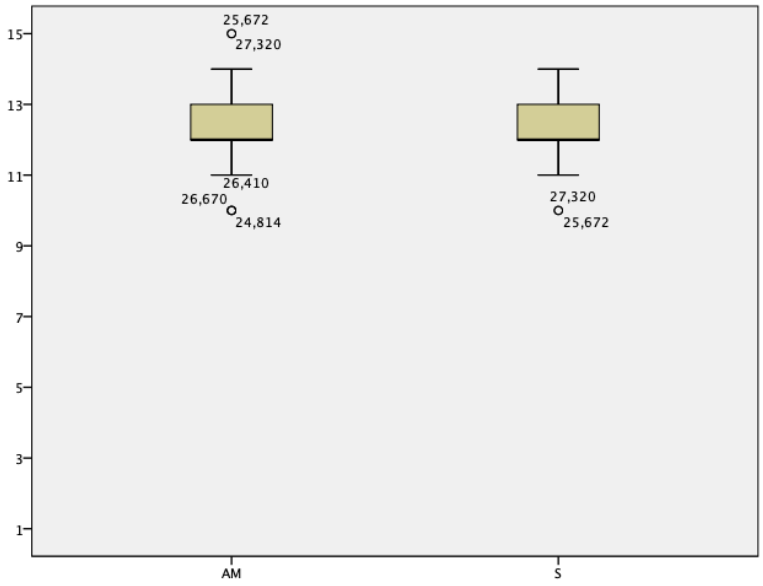


Figure 7
Mean AM and S Scores for Spanish Version D Examinees Rated S



Figures 6 and 7 show very few differences between AM and S scores of examinees rated AM and none for examinees rated Superior. This is corroborated by a paired samples *t*-test (two-tailed) that did not find any statistically significant difference between AM and S mean scores ($t = 1.47$, $p = 0.14$, $df = 149$). Note, however, that only 27 examinees were rated Superior, indicating that there may not have been sufficient cases to establish a reliable relationship. Figure 7

shows that examinees who were rated Superior had most of the AM items correct (median = 12, IQR = 12-13; full range without outliers = 11-14 out of 15).

The data reports in Section 13 showed that there were a number of Superior items who were too easy for both Spanish LPT 01 and LPT 02. In Spanish LPT 01, four Superior items were flagged for revision and an additional four for inspection. In Spanish LPT 02, four Superior items also were flagged for revision and one more for inspection. These revisions have been completed, and it is expected that the revised forms will more clearly distinguish between AM and Superior on the basis of the established cut-scores.

These three sources of evidence, therefore: the results of the side-by-side study; the results of the standard-setting workshop; and to a lesser extent, the analysis of the mean scores of 1,348 administrations of all Spanish two-sublevel tests collectively provide evidence that the cut-scores recommended on the basis of the original side-by-side study are reasonable and appropriate.

Raw Score Conversion to ACTFL Proficiency Levels

LPT raw scores are converted to ACFTL proficiency levels depending on the version of the test, e.g., IL to IM, AL to AM, etc. The same raw scores, therefore, have different meanings depending on the ranges considered. This means that raw scores cannot be used to recommend college credit. Instead, ACTFL proficiency levels may be used. Because ACTFL proficiency levels follow the same logic across the four modalities of speaking, writing, reading, and listening, it is recommended to use the same ACTFL sublevels for listening as are used for speaking and writing. Table 21 shows the recommendations for granting college credit for each ACTFL proficiency level.

Table 26
Recommendations for Granting College Credit

Official ACTFL LPT Rating	**Category I** English, French, Italian, Spanish, Portuguese	**Category II** German	**Category III** Russian	**Category IV** Arabic, Japanese, Korean, Mandarin
Novice High/Intermediate Low	2 LD*	2 LD	3 LD	3 LD
Intermediate Mid	4 LD	4 LD	6 LD	6 LD
Intermediate High/Advanced Low	6 LD	6 LD	8 LD	8 LD
Advanced Mid	8 LD + 2 UD**	8 LD + 3 UD	6 LD + 4 UD	6 LD + 5 UD
Advanced High / Superior	8 LD + 2UD	8 LD + 3 UD	6 LD + 6 UD	6 LD + 6 UD

*LD = Lower division baccalaureate/associate degree category
**UD = Upper division baccalaureate degree category

These recommendations are supported by the results of a nation-wide study examining listening proficiency levels of college students (Tschirner et al., forthcoming; Tschirner & Soneson, forthcoming; see Tschirner, 2016 for similar results). Table 27 shows average speaking and listening proficiency ratings of college students after having completed two, four, six, or eight semesters of studying French, German, or Spanish.

Table 27
Average Speaking and Listening Proficiency Levels of French, German, and Spanish Students at U.S. Colleges and Universities with Numbers of Tests in Parentheses

	French		German		Spanish	
Semester	Speaking	Listening	Speaking	Listening	Speaking	Listening
2	NH (241)	NM-NH (220)			NH (342)	NM (344)
4	IL (284)	NH-IL (265)	IL-IM (194)	NH-IL (312)	IL (436)	NH (418)
6	IM-IH (242)	IM (210)	IM (36)	IH (34)	IM (501)	IM (456)
8	IH (81)	IM (77)	IH-AL (45)	IH (67)	IH (233)	IH (154)

Table 27 shows that speaking and listening proficiency levels of college students are broadly comparable across modalities and languages providing further evidence for the credit recommendations in Table 26 above.

28 Information on Norms and Normative Groups (If Appropriate)

Not applicable

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Bärenfänger, O., & Tschirner, E. (2013). Assessing Evidence of Validity of the ACTFL Listening Proficiency Test (LPT) (Technical Report 2013-US-PUB-2). Leipzig: Institute for Test Research and Test Development.

Impara, J. C., & Plake, B. S. (1997). Standard setting: An alternative approach. Journal of Educational Measurement, 34, 353–366.

Tschirner, E. (2016). Listening and reading proficiency levels of college students. Foreign Language Annals 49, 201-223.

Tschirner, E., & Bärenfänger, O. (2013). Validating the ACFTL Listening Proficiency Test. Poster presented at the 35th Annual Language Testing Research Colloquium (LTRC). Seoul, South Korea, July 1-5.

Tschirner, E., Bärenfänger, O., & Wisniewski, K. (2015). Assessing Evidence of Validity of the ACTFL CEFR Listening and Listening Proficiency Tests (LPT and LPT) Using a Standard-Setting Approach (Technical Report 2015-EU-PUB-2). Leipzig: Leipzig: Institut für Testforschung und Testentwicklung.

Tschirner, E., Gass, S., Hacking, J., Rubio, F., Soneson, D., & Winke, P. (forthcoming). The Role of Listening in the Growth of Speaking Ability.

Tschirner, E. & Soneson, Dan (forthcoming). Acquiring German language abilities in college: Triggers and Dynamics.

Weir, C., & Khalifa, H. (2008). A cognitive processing approach towards defining listening comprehension. Cambridge ESOL Research Notes, 31, 2–10.

Appendices

Appendix 1: Familiarization Manual
Appendix 2: Assessment Use Argument
Appendix 3: Design Statement
Appendix 4: Blueprint
Appendix 5: Construct Matrix
Appendix 6: Technical Report
Appendix 7: French, German, and Spanish LPT Data Reports
Appendix 8: Cut Score Logits and SEMs
Appendix 9: Item Writing Manual
Appendix 10: Item Checklists
Appendix 11: Examinee Handbook
Appendix 12: Certificate